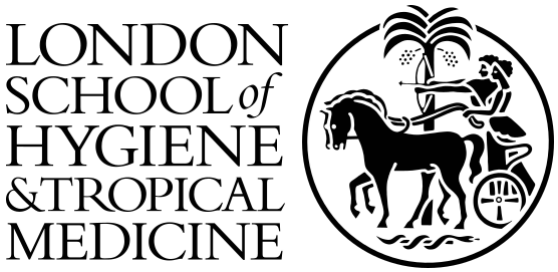
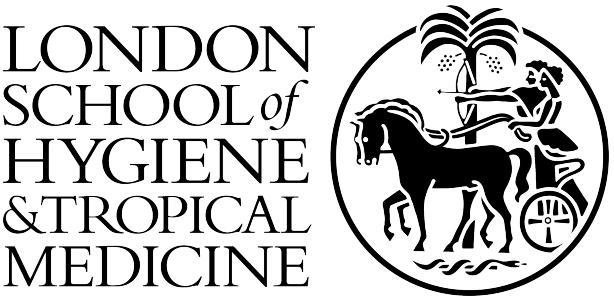




On the prediction and projection of cancer survival

Camille Maringe

Thesis submitted in accordance with the requirements
for the degree of Doctor of Philosophy
of the
University of London

July 2020

Department of Non-Communicable Disease Epidemiology
Faculty of Epidemiology and Population Health
London School of Hygiene & Tropical Medicine

No funding received

Research group affiliation: Cancer Survival Programme

Supervisors

Professor Bernard Rachet

Department of Non-Communicable Disease Epidemiology
Faculty of Epidemiology and Population Health
London School of Hygiene & Tropical Medicine

Dr Aurélien Belot

Department of Non-Communicable Disease Epidemiology
Faculty of Epidemiology and Population Health
London School of Hygiene & Tropical Medicine

Dr Laura M Woods

Department of Non-Communicable Disease Epidemiology
Faculty of Epidemiology and Population Health
London School of Hygiene & Tropical Medicine

Declaration of Authorship

I, Camille Maringe, declare that this thesis titled, 'On the prediction and projection of cancer survival' and the work presented in it are my own.

Signed: [redacted]

Date: 09-02-2020

'All models are wrong, but some are useful.'

George Box

Abstract

Cancer survival is a key metric for monitoring improvement in awareness, early diagnosis and access to effective treatments for cancer patients. For the majority of cancers, survival has been increasing for a number of decades, as a result of successful health policies and the availability of more effective treatment. Nevertheless, there is an unavoidable delay between policy implementation and impact. In parallel, the measure of survival requires follow-up information, adding to the delay in quantifying health benefits. Predictions of cancer survival for cohorts of patients most recently diagnosed could help fill the gap in our knowledge of the likely effects of cancer policies.

In this thesis, I modelled the excess hazard of death as a function of predictors available in linked cancer registry data in the UK. These include age, stage and year of diagnosis, levels of deprivation, type of diagnosis, and access to curative treatment. In such contexts, selecting the form of the model, the predictors, the shape of their effects, and potential interactive effects is challenging. Several model selection strategies are compared and their performance assessed in simulations. I provide practical guidelines for the modelling of the excess hazard of death, in particular in relation to cancer lethality, model complexity and impact of model mis-specification.

Besides, these multi-variable regression models offer opportunities for predicting cancer-related death rate, for cohorts of patients most recently diagnosed, and for whom follow-up is not yet available. Along with model selection algorithms, I explore strategies based on information criteria and model averaging. Inference is therefore conditional on a pool of models of equivalent support, rather than a uniquely selected model. Advantages include absence of multiple testing, and allowance for model selection uncertainty in inference.

Finally, a measure of explained variation, RE, is extended to the relative survival data setting. It is part of the model validation toolkit, and can provide estimates of how much variation in excess mortality due to cancer is explained by the models, and the variables that compose them.

There are several methodological assets from the work presented here. First, excess hazard model selection is well formalised. Furthermore, the way RE is adapted to the relative survival data setting will most certainly nurture ideas for the adaptation of other validation tools, commonly used in prognosis research. Lastly, multi-model inference using model averaging is paving the way for the utilisation of ensemble learning in the prediction of excess hazard of death due to cancer.

Scenario modelling is a public health application that naturally follows the work done in this PhD thesis. With well-crafted set of predictive models, simulated scenarios can be designed to identify areas for improvement in policy, prevention or treatment. Those generating largest increase in survival can lead to actual recommendations.

Methodological advances and public health go hand in hand here. This work emphasises the importance of developing, assessing, and validating excess hazard models. It offers a toolkit so that accurate survival predictions help design effective policies.

Acknowledgements

The PhD journey has been long and sometimes tortuous! Many (many) people along the way, near and far, should be thanked for their continued support. I have learnt as many statistical and epidemiological skills as human skills during these PhD years. And I owe much of this to my three supervisors:

Bernard, you have showed much patience with me! Among many things, you have taught me cancer epidemiology and perseverance. Thank you!

Aurélien, you have taught me statistical rigour and also to recognise and acknowledge little steps forward and small successes along the way. Thank you!

Laura, I will never be as perfectionist as you are, and speak or write English as well as you do, but I have definitely progressed! Thank you!

Thank you all for your availability, your guidance, your smiles and your encouragements!

Thank you to all in the Cancer Survival Group for making the journey more enjoyable! I have had many happy collaborations with Sarah, Miguel-Angel, Nora, Javier, Sara, and Clémence. I have enjoyed working with you on varied topics, and learning from you. This PhD has also been shaped by all these projects running alongside it! Manuela, thank you for walking along this PhD journey at my pace, and for being such a supportive colleague!

There are many friends and colleagues who have made significant contributions by their availability, their eagle eye on the text, their questions, their tap on the back, their support. May they all be thanked here!

My family has been very patient too – if they remember I am registered for a PhD! – and this is an opportunity to thank my parents for their continued support throughout my education.

Christophe thank you for your unfailing faith in me, and for the long evenings you have listened to me!

Timothée, Albéric and Nicolas, I wish your life to be filled with achievements that make you happy! Thank you for your endurance as I was writing this 'book'.

Contents

List of Figures

List of Tables

Abbreviations

AIC	Akaike Information Criterion
AUC	Area Under the receiver operating Curve
BIC	Bayesian Information Criterion
BS	Brier Score
CPD	Crude probability of death
CPRD	Clinical Practice Research Datalink
DCI	Death Certificate Initiated
DCO	Death Certificate Only
DID	Diagnostic Imaging Dataset
HES	Hospital Episode Statistics
IARC	International Association for Research on Cancer
IC	Information Criterion
ICBP	International Cancer Benchmarking Partnership
ICD	International Classification of Diseases
KL	Kullback-Liebler
LASSO	Least Absolute Shrinkage and Selection Operator
MDT	Multi-Disciplinary Team
NBOCA	National Bowel Cancer Audit
NHS	National Health Service
pABCtime	Proportion Area Between Curves through time
PE	Prediction Error
PP	Pohar-Perme
PROGRESS	Prognosis research strategy
RE	Rank Explained
RMISD	Restricted Mean Integrated Square Differences

ROC	Receiver Operating Curve
RS	Royston and Sauerbrei
SACT	Systemic Anti-Cancer Therapy
Se	Sensitivity
SEER	Surveillance Epidemiology and End Results
Sp	Specificity
STraTOS	STRengthening Analytical Thinking for Observational Studies
WA	Wynant and Abrahamowicz

Dedicated to all men and women,
boys and girls, with no access to education.

Glossary

Background (population, expected) mortality: levels of mortality that would be expected in the cancer patient population, if cancer was not affecting mortality patterns. Background mortality is estimated based on what is observed in the general population from which cancer patients come from (from population life tables).

Cause-specific setting: data setting in which cause of death is available.

Censoring: censoring happens when there is loss to follow-up (see below) at a given date before the outcome can be observed. It could be administrative censoring which happens at a set date, when the survival time of patients still alive (in the risk set) is truncated and their vital status is 'alive'. Other types and times of censoring reflects emigration, or losses of information.

Competing risks: death may happen as a consequence of various causes. Cancer survival analyses correspond to estimating survival from cancer; other causes of death act as competing risks, and censor patients' survival time.

Excess (cancer) mortality: mortality in excess of the population levels of mortality (due to cancer).

Life tables: tables of mortality rates defined for the population from which cancer patients are drawn. They are defined by calendar year, single years of age, and sex at minima, and other factors that can be available in the population file and cancer patient data.

Loss to follow-up: this happens when a patient's record is incomplete, that is when the actual length of survival and survival status are unknown.

Net survival: survival from cancer in the hypothetical situation in which cancer patients are immune to other causes of death.

Non-parametric: there is no shape imposed on the relationship between explanatory factors and outcomes.

Parametric: parameters are estimated from the data to define the relationship between explanatory factors and outcomes.

Prediction: estimation of cancer patient outcomes further away from their diagnosis than the data available for model-building.

In-sample predictions: predictions of cancer outcomes for values of variables who are available in the sample of patients on whom the effects are modelled.

Out-of-sample predictions: Predictions of cancer outcomes for values of variable not observed in the sample of patients on whom the effects are modelled. For instance it could be for patients diagnosed in an age group that is not represented at all in the data used for model-building.

	Range observed	**Predictions**	
Variables available:		**in-sample**	**out-of-sample**
Age at diagnosis (years)	15-85	15-85	<15 or >90
Year of diagnosis	2005-2010	2005-2010	2011+
Stage at diagnosis	I-IV	I-IV	NA
Deprivation quintile	1 to 5	1 to 5	NA
Follow up (years)	0-6 years after diagnosis	0-6 years after diagnosis	>6 years after diagnosis

Projection: estimation of cancer patient outcomes for patients that did not contribute to model building. This is a specific case of out-of-sample prediction, in relation to year of diagnosis.

				Follow up			Model-based predictions				
Cohorts	2005	2006	2007	2008	2009	2010	2011				
2005	0/1	1/2	2/3	3/4	4/5	5/6					
2006		0/1	1/2	2/3	3/4	4/5	5/6				
2007			0/1	1/2	2/3	3/4	4/5	5/6			
2008				0/1	1/2	2/3	3/4	4/5	5/6		
2009					0/1	1/2	2/3	3/4	4/5	4/5	
2010						0/1	1/2	2/3	3/4	3/4	4/5
2011		Model-based projections					0/1	1/2	2/3	3/4	4/5

Relative survival setting: data setting in which cause of death is unavailable.

Introduction

Rationale

Long-standing population-based cancer registries operate in many countries in the world. [1] Their resources may vary depending on the local context, but they usually collect a minimal well-standardised set of data in order to provide accurate estimates of cancer incidence in a clearly defined population. Information on follow-up and date of death are collected in order to estimate survival.

The quantity of data it is now possible to collect, store and use has increased dramatically. In England, beyond the data collected by cancer registries, detailed information on patient, tumour, and management for cancer patients can be obtained from primary care (such as in the Clinical Practice Research Datalink), secondary care (such as in the Hospital Episode Statistics data), clinical audits (such as National Bowel Cancer Audit, and the Lung Cancer Audit) and Multi-Disciplinary Teams, administrative data sources such as insurance claims and imaging or pathological laboratory information (such as the Systemic Anti-Cancer Therapy or Diagnostic Imaging Dataset). This increasingly detailed amount of information can help deepen the understanding of complex pathways to diagnosis, disease progression, clinical management, and ultimately, survival.

Cancer survival estimates are calculated at a given time t, most often expressed in years. Cancer survival measures the probability that patients with a given cancer will survive beyond t, derived from the observed proportion of patients who do survive beyond t. These estimates are of interest to a wide range of actors including patients, clinicians, insurance companies, policy-makers, and the public. All parties need these figures for different purposes, and the presentation of these estimates sometimes needs to be adapted to their requirements for optimal interpretability. [2, 3]

For the benefit of the cancer patients as a whole, a better comprehension of the mechanisms driving the levels of and trends in cancer survival is essential. Such understanding is mainly derived from statistical models aimed at representing complex interactions in a

simple framework. These models are the basis for population-based outcome predictions. Cancer survival is a key measure for monitoring the impact of health policies in the population. It can be derived from modelling the effects of predictors on cancer mortality. These predictors are collected primarily by cancer registries, but estimating the effects of additional, potentially prognostic variables provides further insights into the mechanisms underlying patterns of survival. Statistical tools are available for complex modelling of the effect of individual factors on cancer mortality, as well as user-friendly statistical packages. [4–9]

The Calman-Hine report, published in 1995, [10] set out to improve outcomes and reduce inequalities in cancer care. Since then, the succession of health plans in England shows a focus on improving the performance and equity of the health care system. Studies that analysed the impact of the first comprehensive policy on cancer services were mostly published a decade later, [11–13] highlighting the unavoidable time-lag between implementation and assessment. In that timeframe, the NHS Cancer Plan had been introduced in 2000 [14] and further policy documents followed. [15, 16] Health-policy makers are eager to see the impact and evaluate the effectiveness of newly introduced strategies. Unfortunately, the nature of survival itself means that it takes many years to gather follow-up information in order to produce meaningful and reliable estimates. Although guidelines are often implemented based on an estimated predicted impact, their perceived and genuine effects remain uncertain until they can be evaluated. Furthermore the real-world effects of policies are impacted by the health system complexity. Such effects are often unpredictable before implementation.[17] Prediction of survival is aimed at answering these questions in a more timely fashion. By prediction of survival, we mean the estimation of survival for patients only very recently diagnosed, or the estimation of long-term survival for patients who were diagnosed some years ago but for whom we do not yet have information on their longer-term follow-up.

Thus, there is a need to consider and develop methodology for the prediction of cancer survival, at a population level. The main goal of this thesis is to progress the methodology for predictions and projections of cancer survival for the population as a whole, in contrast to individual prediction models. In this work, the unit of prediction is the population. The population may include only patients with given characteristics (such as a given stage, or age group) or may be an entire cohort of patients diagnosed in a given time and place. In contrast with prediction, projection of cancer survival refers to estimates for cohorts of patients who have yet to be diagnosed. These cohorts can be fully hypothetical (scenarios), or refer to patients who have already been diagnosed, but for whom no follow-up information is yet available.

Statistical models utilise patient data to estimate the average effects of predictors. Outcomes, such as the mortality hazard, are predicted for individual patients, but can also be transformed and averaged to provide population-based predictions and projections of net survival, marginal effects and crude mortality.

This work stands at the cross road between three different fields: overall prognosis, prognostic factors and prognostic models, as highlighted in the PROGRESS framework. [18–20] The focus of 'overall prognosis' research is in the description of actual observed outcomes. The identification of prognostic factors influencing outcomes is next. Prognostic modelling is concerned with individual predictions from these models, built on sample of patients. There is much emphasis on developing individual prediction models in a context in which targeted therapies and individual treatment are being offered. Several authors provide guidelines for adequate development and validation of prognostic models to guarantee they are appropriate and useful to guide the clinical and patient decision-making process. [18–23]

Due to the population-based nature of this work, we are concerned with overall prognosis and average outcomes, for a population. We claim such outcomes are best estimated via multi-variable models. These models are also key to make projections for groups of patients for whom follow up information is not yet available. Therefore, we borrow tools from all three fields in order to propose a methodology suitable to our context of population-based predictions of cancer survival. Such tools include parametric modelling and algorithms for model selection and explained variation. Additionally we evaluate predictions from multi-model inference based on model selection using information criteria. We judge the quality of predictions for groups of patients defined by their shared characteristics.

Aims

1 To provide guidelines for the selection of multivariable excess hazard models in the context of descriptive epidemiology.

2 To contribute to research in statistical methods for the evaluation and validation of multivariable predictive excess hazard models.

3 To investigate the use of multi-model inference for the prediction and projection of cancer survival based on multivariable excess hazard models.

Objectives

1a Describe model selection strategies in the context of excess hazard models and exemplify their use in the estimation of cancer survival.

1b Perform model-selection for the estimation of cancer survival from multivariable excess hazard models.

1c Conduct multivariable exploratory analyses of the effect of predictors such as age, deprivation, stage, treatment, and screening, where available, on the estimation of cancer survival.

2a Adapt a measure of explained variation to the context of excess hazard regression models.

2b Apply measures of explained variation to multivariable excess hazard models.

3a Develop excess hazard model selection based on Information Criteria.

3b Implement model averaging techniques for the estimation of excess hazard.

3c Predict survival for patients recently diagnosed using multi-model inference.

Setting: Estimating cancer survival using population-based registry data

1 Introduction

The aim of this chapter is to introduce the specific context of modelling cancer survival from observational population-based cancer data. Opportunities and challenges of this setting are discussed, for example the large amount of population-based (+) data available (+), lack of patients' selection (+/-), missing information (-), and presence of competing risks (-).

I use data on *cancer patients* – the unit of information is any man or woman with a diagnosis of cancer; collected in *registries* – routine collection of information on patients; at *population level* – the conclusions from the analyses will apply to all people living in a defined territory.

2 Population-based cancer surveillance

2.1 Cancer registration

Many countries benefit from cancer registration systems, at regional or national level. Population-based cancer registries cover the population of a clearly defined geographical area and register all incident cases of cancer that arise within it. A minimum set of variables is required so that age-specific cancer incidence patterns can be studied. These include patient (date of birth, sex and a unique identification number) and tumour factors (date of diagnosis, cancer site, behaviour code and a unique identification number). High-quality, un-interrupted registration enables valid and accurate surveillance.

In the UK, survival is among the key metrics which can be estimated using registry data. Cancer registries ordinarily obtain information on each patients' vital status through linkage with national mortality records. Additional variables may also be collected and allow detailed analyses of cancer incidence and survival by patient socio-demographic (for example, sub-region, place of birth, ethnicity) or tumour (for example, grade, laterality, stage at diagnosis) characteristics. When the registry collects clinical information or if the data sets are linked to specialised clinical registration systems, further analyses evaluating the impact of treatment can be performed.

Cancer registries use the International Classification of Diseases [24] to register and code tumours to a specific anatomical sub-site. The tumour is classified according to how it looks under a microscope: its type of cells (morphology) and whether its behaviour is benign, in-situ or malignant. Together these characteristics can be used to define groups of patients with similar disease. The proportion of patients with microscopic verification within a population-based registry tends to be reported as a marker of the quality and thoroughness of the registration system. Ordinarily only patients with malignant cells are analysed. Nonetheless, the proportions of patients diagnosed with benign/in-situ disease may be reported as an indication of diagnostic intensity and coverage of the cancer registry.

2.2 Date of diagnosis

The date of diagnosis is a key variable in cancer registration. Firstly, it provides the context around the cancer diagnosis. This is also known as the cohort effect, which regroups the effects of differences in intensity or availability of tools for diagnostic investigations, differential availability of screening tests, varying coding practices, etc. Secondly, it determines which cohort a patient belongs, and for whom specific characteristics may be explored, and survival may be estimated. Lastly, survival analyses can only be done when time zero is clearly specified, marking the start of follow-up. Date of diagnosis is unknown for patients whose record is initiated from a death record mentioning cancer and for whom no further information is ever found; these patients cannot be included in survival analyses.

Two standards are established for the registration of the date of diagnosis: (a) IARC (International Association for Research on Cancer) working group [25] and (b) SEER (Surveillance Epidemiology and End Results) rules. [26] The date of diagnosis following the IARC rules is the date of the histological diagnosis or the actual date (not result) of biopsy. SEER rules take first date between clinical assessment or the date the histological report is issued. For both sets of rules they recommend using the treatment date if a patient receives it before definitive diagnosis or if it leads to a diagnosis.

2.3 Cancer incidence, survival and mortality

This thesis focuses on the prediction of cancer *survival*. However, survival, incidence and mortality are inter-dependent and must be considered alongside each other (see Figure 1). The study of survival cannot be done in isolation and must be understood in the context of fluctuating incidence and mortality patterns.

Both cancer incidence and mortality are functions of the characteristics of the general population: cancer incidence reports the number of new cases of cancer and cancer mortality reports the proportion of people who die from cancer, in a population of a specific time and place. Although they refer to a given (possibly current) year, they result from past exposures to carcinogens, past and present participation in screening practices (incidence) or effectiveness of treatment and prompt cancer diagnosis (mortality). In contrast, cancer survival reports the proportion of patients diagnosed in a well-defined time and place, who survived their disease at given milestone dates, commonly 1, 5 or 10 years after diagnosis.

The combination of all three measures represent a comprehensive picture of the cancer burden. [27] They inform on the management of cancer, by studying its causes and outcomes, and how these change through time or are affected by varying public health policies. Thus, the study of survival cannot be performed in isolation and must be understood in the context of fluctuating incidence and mortality patterns.

Since survival is intertwined with incidence and mortality, future trends in survival must be considered in the context of current and likely future levels of incidence and mortality. For *prediction* of cancer survival, we assume the composition of cohorts of patients is known. It means patients have already been diagnosed with the disease, possibly recently, and their diagnosis, along with socio-demographic and clinical information, are documented in the cancer registration data. Due to their recent diagnosis and lack of follow-up information, only little is known on their survival pattern, and certainly nothing is known on their long-term prognosis. Prediction of survival refers to the estimation of the proportion of patients who survive until a given time, further away from diagnosis than what has been observed. Incidence is therefore known, or controlled, in this context.

In the context of *projection* of cancer survival, the actual cancer incidence, i.e. the cohorts of cancer patients diagnosed at a future date, may or may not have been observed yet and their survival is similarly unknown. When incidence is not yet known or recorded, different scenarios can be envisaged:

(a) Changes in incidence which are the results of public health policies: These would influence the overall distribution of patient and/or tumour characteristics, but the effects of the predictors on survival would remain unchanged.

(b) Changes in the effects of predictors on cancer survival: We assume the composition of the cohorts remains unchanged.

2.4 Follow-up information

The analysis of cancer survival relies on the ability of cancer registration systems to follow-up patients, in order to assess their vital status after diagnosis or regularly link their cancer records with national death records. The date of last known vital status needs to be available to derive the length of follow-up for every patient. Follow-up of cancer patients is achieved in one of two different ways: passive or active.

In a passive follow-up setting, patients are assumed to be alive until evidence to the contrary is acquired, normally when a routinely collected death record is linked to their cancer record. Patients' follow-up time can also be censored if they make the national statistics office aware of their emigration. The caveats of such a follow-up assessment is that some patients may become 'immortals', when their emigration status is not known, or where there are linkage inaccuracies. Linkage to additional datasets, such as general practice registration data minimises such issues. Additionally, cancer registries can receive information from a death with underlying cause flagged as cancer but for whom no tumour record is (yet) held on the system. This record is termed a death certificate initiated (DCI) registration. If no further information on the cancer is ever found, it turns into a death certificate only (DCO) record because no other information on the tumour is held in the registration system. DCOs must be flagged so they do not contribute to survival analyses since their follow-up time is effectively unknown.

Active follow-up is when researchers actively seek information on the vital status of patients in their registry, through for example primary or secondary healthcare providers, or via the city of residence using direct correspondence to assess the vital status of each patient. This second follow-up technique is very labour intensive, and is typically done only for small datasets, in settings with poor national registration systems or when linkage between administrative datasets is prohibited.

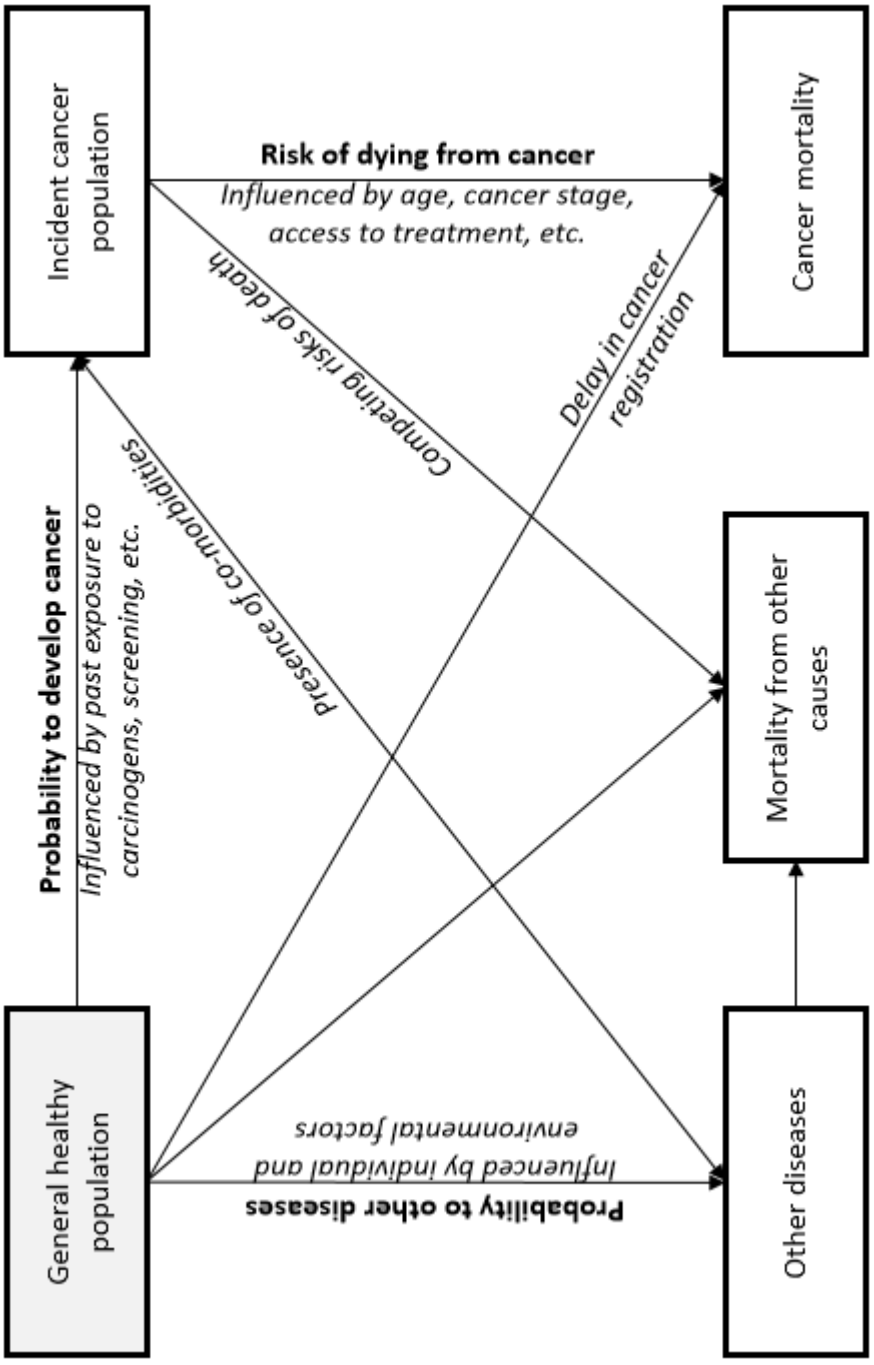


Figure 1: Factors at play in the variations in incidence, mortality and survival of cancer.

2.5 Data preparation for survival analyses

Following detailed quality control rules for assessing cancer registration records defined by IARC, patients with incomplete key information or incongruous records such as incomplete dates or sex/site contradictions are first excluded from the data. [28, 29] These procedures ascertain that incomplete, incoherent and ineligible records are excluded. The necessary additional information on length of follow-up for survival analysis means further data quality checks are done before one can produce cancer survival figures. [30] These include checking records for invalid sequences of dates, unknown vital status, or DCO registrations which cannot be utilised in survival analysis since exact survival time between cancer diagnosis and end of follow-up is not known.

3 Estimation of cancer survival

3.1 Different measures for different purposes

Population-based cohorts of cancer patients offer an opportunity to estimate the cancer burden in the population, and look at its trends over time. Both the motives for performing survival analyses and the end users of the estimates (policy-makers, patients, or clinicians) determine which measure is most appropriate. [2]

Several questions of interest can be asked and each measure answer a specific question: [2]

- *What is the survival of cancer patients?*

  This would be best addressed by the estimation of *overall survival*: the probability that cancer patients live beyond pre-defined milestones.

  Overall survival is estimated non-parametrically, using Kaplan Meier survival curves, [31] semi-parametrically, with the Cox model [32] or fully parametrically, using standard distributions such as Weibull.

- *What is the probability that cancer patients die of their cancer? Or of other causes?*

  This is reflected in *crude mortality*: the proportions of deaths, in cancer patients, that are due to cancer, accounting for the competing risks of other causes, or due to other causes of death accounting for the competing risks of death from cancer.

- *What is the cancer survival of cancer patients?*

  Here the interest is in the proportion of patients who survive their cancer. To measure the impact of cancer alone on patients' survival, we need to remove the effect that competing risks have on survival. This is *net survival*.

Both crude mortality and net survival can be estimated with the cancer registration data, the next sections highlight how this can be best done.

3.2 Competing risks

Within cancer registry, routine death certification is the source of information for the date and cause of death of patients who have died. Patients may die of their cancer, or of causes other than their cancer, such as consequences of their cancer treatment, or any other cause possibly unrelated to cancer. When studying *cancer survival*, we are usually most interested in survival from cancer alone and all other possible causes of death thus act as competing risks for cancer death. Cancer survival can be estimated through net survival, assuming patients can only die of cancer. In reality, it is impossible to observe a cancer survival time for patients who die of causes other than cancer, as the competing event is preventing the observation of the event of interest. Two complications follow:

1. How do we assess if the event of interest was observed or not?
2. How do we include information from patients whose event of interest cannot be observed because of competing risks such as deaths due to other causes?

In randomised clinical trials, the first question is addressed by strict protocols to standardise the collection and recording of patient-related information, including the identification of their cause of death. By doing so, uncertainty about what constitutes an event of interest is minimised. In routine death certification, there are international rules for recording and coding cause(s) of death on death certificates applied globally. Nonetheless, there are still wide temporal, geographical and inter-personal variability in the routine registration of cause of death. This means that over time and between regions, cause of death as reported on the death certificates may not be comparable. More importantly, it may also be difficult to separate deaths that have occurred as a consequence of cancer, directly or indirectly. For these reasons, within population-based analyses of cancer survival, it is challenging to assess whether the event of interest is observed or not.

To answer the second question, one needs to adjust the analyses for the imbalance between the 'surviving' cohort and the cohort of patients entering the analyses at diagnosis (time zero). Patients who die throughout the follow-up, of causes other than cancer, are not necessarily similar to other patients in the cohort. Indeed, patients who survive all competing risks of death are more likely to be younger, fitter, have less co-morbidities and less aggressive tumours. The information brought by patients who eventually do not survive competing risks of death needs to be accounted for in the calculation of cancer

survival since the characteristics associated with that censoring (mortality from competing risks) are also associated with cancer survival, and as such, censoring is informative. [33] The methods presented below (section 3.5) account for informative censoring and provide unbiased estimates.

3.3 Net survival and excess hazard

Net survival refers to the cancer survival that a cohort of cancer patients experience. The cohort of patients may be defined based on when patients were diagnosed (calendar time), where they were diagnosed (region), their sex, age at diagnosis, ethnicity, or tumour characteristics such as site (anatomical), type (cancer cells) or stage at diagnosis. Cancer survival measured for the cohort reflects what the survival of cancer patients would be, if cancer patients could only die of their cancer. The measure of net survival is derived purposefully to isolate the effect of cancer on survival, over and above competing risks. It is not observed in practice. Free from the impact of competing risks, net survival enables fair comparisons of survival for different groups of cancer patients, without being affected by differences in mortality due to other causes.

Net survival for the cohort is a cumulative measure estimated at all times after diagnosis as long as there is information for that time point, that is, patients alive as well as events occurring (deaths). By contrast the excess hazard is the cancer mortality hazard, measured at a given time t. It is an instantaneous measure calculated for each individual patient in the cohort. It represents the likelihood for the patient to die *of their cancer* at time t.

The measures described here reflect *population* cancer survival and *individual* cancer mortality respectively. Their accurate estimation relies on the separation of the risks of death competing with the risk of dying from cancer.

3.4 Data settings

Depending on the type of information available, there are two different data settings: (a) the cause-specific setting, in which reliable information on the cause of death is available, and (b) the relative survival setting, in which cause of death information is not reliably available. Net survival and excess hazard of death from cancer can be estimated in both data settings.

The context of population-based cancer registry data is the perfect example of a *relative survival data setting* in which, in order to estimate cancer survival, the competing risks of death (i.e. of the causes of death other than cancer) are most commonly estimated using

risks of death derived from the population from which the cancer patients are drawn. It is also possible to use the information contained in the underlying cause of death available on death certificates in order to attempt to tease out which are deaths due to cancer. This corresponds to the estimation of net survival in a *cause-specific data setting* framework. [34, 35] Quality, accuracy and reliability of cause of death is paramount for the estimation of net survival in the cause specific setting. [36]

3.5 Life tables

Definition

In the relative survival data setting, *life tables* are crucial to the estimation of net survival. Population life tables are estimated from the general population from which cancer patients are drawn. They provide estimates of all-cause mortality for each patient at the end of their follow-up (Figure 2). This adjustment must be made in order to isolate the cancer-specific excess mortality and derive an estimation of net survival.

At the time of their last known vital status, demographic information of the patients is used to determine their expected mortality rates. These expected population mortality rates are used in survival models to adjust the observed (all-cause, overall, λ_O)) mortality for mortality due to other causes (non-cancer, background, expected, population, λ_P). The most common assumption is that all-cause mortality is the sum of the mortality due to cancer (excess, λ_E) and due to other causes for each patient i. This is the additive model: $\lambda_{Oi}(t) = \lambda_{Ei}(t) + \lambda_{Pi}(t)$.

Population tables of mortality rates, or risks of dying in yearly intervals, life tables, are defined by sex, single years of age, and calendar year, at minima. They are the basis for the calculation of life expectancy at birth. They represent patterns of mortality for all living in a defined geographical area. Life expectancy at birth varies between regions of a country, but also between deprivation groups and/or ethnic groups. [37] When such detailed information on deprivation levels and ethnicity is available on both population and death counts, as well as on the cancer registry data, life tables further defined according to the levels of these socio-demographic variables can be generated. [37–40] This allows taking those variations into account in the estimation of the background mortality, and not attribute them to variations in cancer care and survival (Figure 2). In multivariable modelling of cancer survival the effect of socio-economic deprivation is estimated, and as such, adjusting for all-cause mortality stratified by deprivation quintile is relevant.

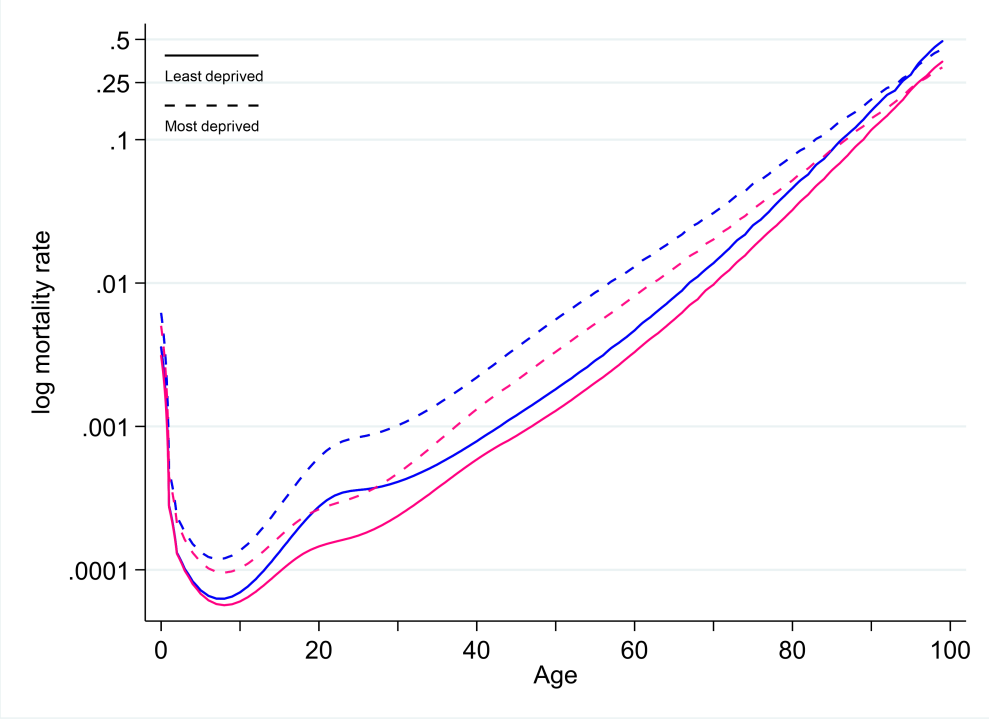


Figure 2: Estimates of mortality rates, by single year of age, and sex (male – blue, female – pink) and for most (dashes) and least (plain) deprived quintiles of the English population in 2011

Constructing life table

Life tables can be constructed where unavailable. [38, 39, 41] This can be done from raw counts of deaths and population, ideally by single year of age, and any other variable that strongly influences population mortality. It may be that background mortality is 'mismatched' to that of the patients, i.e. it does not correspond to the true force of mortality that the patient would experience in the absence of cancer. The mismatch can reflect a lack of stratification of the background mortality, or that the cancer patients have specific characteristics that differentiate them from the general population. Depending on the direction of the mismatch (under or over-estimation of background mortality), it leads to a biased estimation of excess mortality. For instance, socio-economic differences in cancer survival are increased when variations in background mortality by social deprivation are not considered.

After calculating raw mortality rates, we may get variability due to scarce number of events in small groups of the population, in each calendar year and age. Various methods are available to introduce stability and one can smooth mortality rates so that mortality patterns are not fluctuating too much between ages or calendar years. [42]

3.6 Statistical methods for estimating net survival

As already stated, net survival is not directly observable: it is the survival that would be observed if patients could only die from their cancer. In addition to the notations of the three types of hazard (λ_O, λ_P, λ_E) given above, we define the following notations:

$S_O(t)$ is the overall survival for the cohort (all patients), at time t. It is an estimate of the proportion of patients still alive at t.
$S_E(t)$ is the net survival for the cohort at t, and $S_{Ei}(t)$ is the individual net survival value for patient i at time t. It is the probability that patient i survives their disease longer than time t, given their prognostic factors X_i : $S_{Ei}(t)$=$p(T_{Ei} > t|X_i)$. T_{Ei} is patient i cancer related survival time.
The relation that links hazard and survival, namely $S_O(t) = \exp\left(-\int_0^t \lambda_O(u)du\right)$ in classical survival analyses, remains valid in the excess hazard setting, for each patient i: $S_{Ei}(t) = \exp\left(-\int_0^t \lambda_{Ei}(u)du\right)$.
$\lambda_{Ei}(t)$ is the individual excess hazard of death for patient i at time t.

At any time t, the net survival of a group of N patients is the average of the individual net survival probabilities for patients in that group:

$S_E(t) = \frac{1}{N}\sum_{i=1}^{N} S_{Ei}(t)$.

Estimating net survival: Non-parametric approach

Similar to the Kaplan Meier estimation of overall survival, [31] there is a non-parametric estimator of net survival. [33] At each event time, the net survival for each patient i in the cohort is estimated using their observed mortality in relation to their expected mortality as determined by the life tables.

The non-parametric Pohar-Perme estimator of individual cumulative excess hazard (and population net survival) is the estimator of choice. [33] It accounts for informative censoring relative to the presence of competing risks of death (i.e. withdrawal of patients from the cohort, due to other causes), using the inverse probability of censoring as weights. Indeed, each patient's contributions to the overall measure of survival is weighted by the inverse of their individual expected survival probabilities derived from population life tables, S_{Pi}.

From the additive model assumption, we have $\lambda_{Ei}(t) = \lambda_{Oi}(t) - \lambda_{Pi}(t)$. $\lambda_{Oi}(t)$ is estimated as the ratio of the number of observed events $dN(t)$, in a small interval dt, over the number of patients at risk at the start of the interval, $Y(t)$. We weight all of its components: $\lambda_O(t) = \frac{dN^w(t)}{Y^w(t)}$ with $Y^w(t) = \sum_{i=1}^{N} \frac{Y_i(t)}{S_{Pi}(t)}$ and $dN^w(t) = \sum_{i=1}^{N} \frac{dN_i(t)}{S_{Pi}(t)}$.

Similarly, the population hazards λ_{Pi}, estimated from population life tables, is corrected for informative censoring as follows: $\lambda_P(t) = \frac{\sum_{i=1}^{N} Y_i^w(t)\lambda_{Pi}(t)\,dt}{Y^w(t)}$.

The Pohar-Perme estimator estimates the cumulative excess hazard of death and is defined by:

$$\hat{\Lambda}_{Ei}(t) = \int_0^t \frac{dN^w(u)}{Y^w(u)} du - \int_0^t \frac{\sum_{i=1}^{N} Y_i^w(u) d\Lambda_{Pi}(u)}{Y^w(u)} du. \quad (1)$$

Net survival for a given patient i is such that $\hat{S}_{Ei}(t) = exp(-\hat{\Lambda}_{Ei}(t))$. We estimate net survival for the entire cohort of patients by taking the average of each individual's value: $\hat{S}_E(t) = \frac{1}{N_t} \sum_{i=1}^{N_t} exp(-\hat{\Lambda}_{Ei}(t))$. N_t is the number of patients in the risk set at time t.

Net survival can be estimated for the entire cohort of patients, or for more homogeneous groups of patients sharing specific characteristic(s) such as the type or severity (stage) of disease with which they were diagnosed, the age at which they were diagnosed, their ethnicity, and so on and so forth. One aspect of non-parametric estimation is that the effects that such predictors may have on cancer survival or mortality cannot be estimated, since their values stratify the analyses.

Furthermore, stratifying may lead to sparse data and estimates with large variance due to lack of information. As with all non-parametric estimators, the measure of net survival is sensitive to the sparsity of events in the data. As the number of events decrease, each event that occurs carries a lot of weight as each represents a larger share of the data, and therefore each event has a larger influence on the estimates. Lastly, with the correction for informative censoring using inverse probability of censoring weights (probability to survive beyond time t), each event is weighted to represent patients in the cohort who died of causes other than cancer: the higher the probability of dying of other causes, the larger the weights (for example amongst older patients). [43]

Estimating net survival: Regression models

In contrast to non-parametric estimation, excess hazard regression models estimate the effects of individual predictors, such as age at diagnosis, sex, ethnic group, ecological or personal deprivation (possibly stratifying the life tables), and tumour factors such as stage at diagnosis, on the excess hazard of death.

Like non-parametric estimation, excess hazard regression models adjust for background mortality by linking estimates of all-cause mortality (life tables) to the patients' records at their time of death or censoring.

Model-based individual net survival estimates, $\hat{S}_{Ei}(t)$ are obtained for all patients at each time t throughout follow-up regardless their at-risk status and possible censoring. Overall net survival for the cohort is a simple average of these individual estimates, at each time t. This is fundamentally different from the non-parametric approach in which $\hat{S}_{Ei}(t)$ could only be estimated for patients alive at time t.

In an excess hazard model, the excess hazard of death is a function of the effect of time since diagnosis t, as well as of the effects of potential prognostic factors X. The most common form for the excess hazard model is multiplicative:

$\lambda_E(t) = \lambda_{0,E}(t) * exp(f(X, t, \beta))$

$\lambda_{0,E}(t)$ is the baseline excess hazard that is estimated at each time since diagnosis, and at the reference values of all predictors. f can be a flexible function of time t, X, vector of predictors and β their corresponding parameters. Different parameterisations exist for the modelling of the baseline excess hazard and for the relationship between predictors and excess hazard.

Modelling the baseline excess hazard of death

The baseline excess hazard represents the change in cancer-related mortality through follow-up time, at the reference values of the predictors. One can impose a standard distribution (such as Poisson, Weibull, log-logistic) on the baseline excess hazard ($\lambda_{0,E}$). The parameters defining the standard distributions are estimated by maximum likelihood using the data. Alternatively, for more flexibility in the baseline excess hazard, one can estimate the parameters of fully flexible functions entirely derived from the observed data points, such as fractional polynomials, [44] restricted cubic splines, [45–47] B-splines [6] or penalised tensor splines. [5] Fluctuation of the baseline excess hazard with time needs to be modelled with care.

Modelling continuous and time-dependent effects on the excess hazard

Let us take the example of the effect of age at diagnosis on lung cancer mortality to illustrate the choices available for the parameterisation of the effect of a continuous variable on excess hazard. Figure 3(a)-3(c) illustrates different shapes for the effect of age at diagnosis on cancer mortality based on different parameterisations of age:

- A *linear effect of age* – a linear effect implies that one parameter, $exp(\beta)$ is enough to characterise the relationship between age and cancer mortality, for all ages and at all times: in Figure 3(a), there is a 1.2-fold increase in excess mortality with a 10-year increase in age.
- A *non-linear effect of age* – a non-linear effect tries to capture a more complex relationship between age and mortality. Depending on the complexity of the association between age and mortality, it may be monotonically increasing or decreasing, or have a bell shape. In Figure 3(b), the effect of aging on mortality for ages below 50 is reduced, but increased for ages above 65: the hazard ratio for a 10-year increase between ages 72 and 82 is increased to 1.3.

- A *time-dependent effect* – a time-dependent effect is simply an interaction between the main effect (age at diagnosis for example) and follow-up time. In Figure 3(c), an additional time-dependent effect reveals how the effect of age at diagnosis, at all ages, is stronger 1 month after diagnosis, but milder 12 months after diagnosis. Time-dependent effects may be estimated for continuous and categorical variables alike. When dummy variables are considered for the modelling of a categorical variable, either each individual dummy variable or all of them can be treated as time dependent.

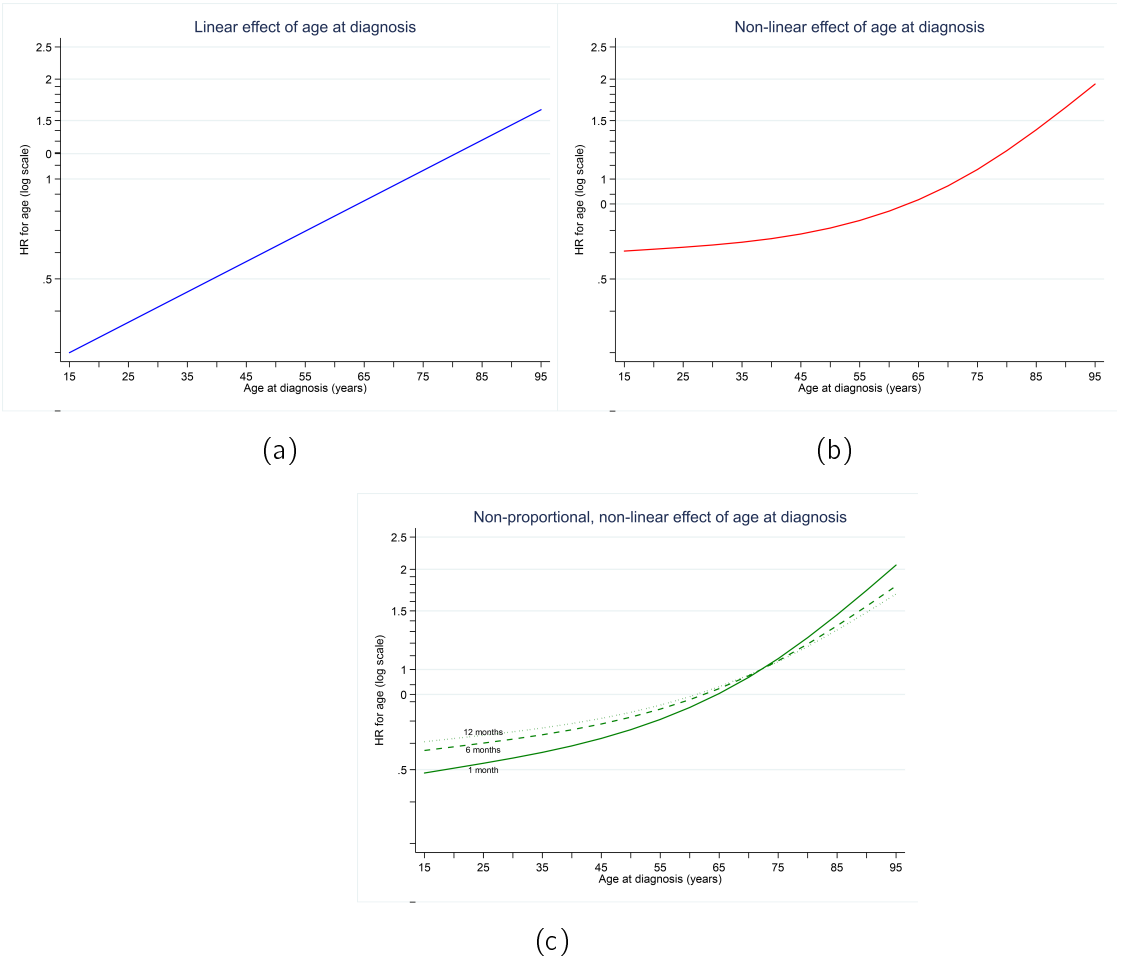
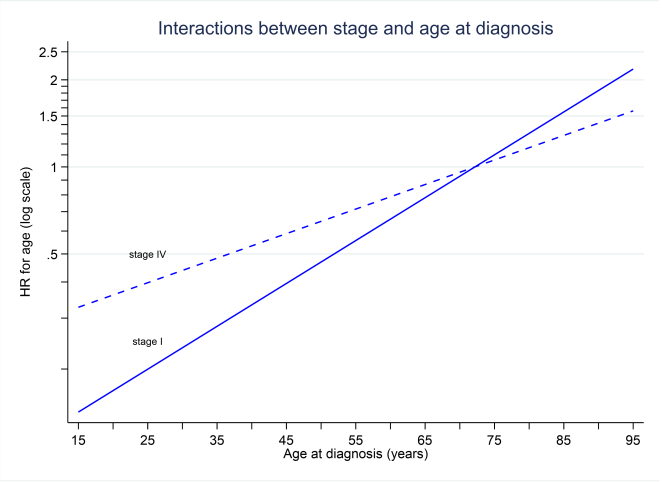


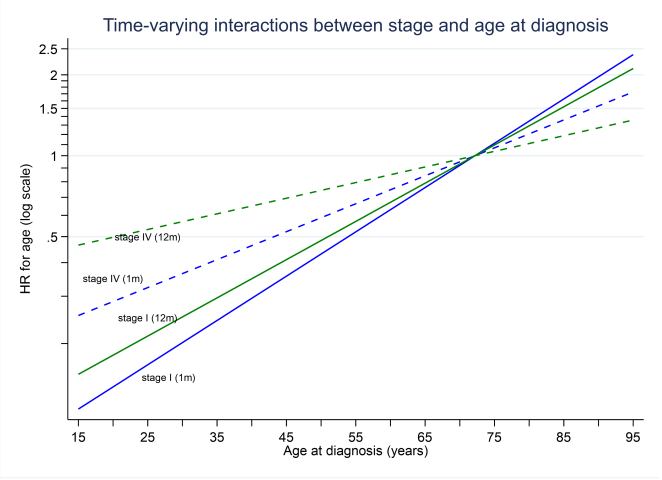
Figure 3: Varying shapes for the age effect, based on varying assumptions

Modelling interactions between covariates

An interaction is modelled when there is evidence that the effect of a variable on excess mortality is modified by the values of another variable. Time-dependent effects are special case of interactions whereby the effects of a variable are modified through follow-up time. It is common practice to create dummy variables for interactions, ahead of model fitting. Interactive effects can be fixed in time (Figure 4(a)) or time-dependent, in addition to their main effects (Figure 4(b)).



(a)



(b)

Figure 4: Interactions between the effects of age and stage at diagnosis

3.7 Assumptions

When estimating cancer survival from population-based cancer registry data in the relative survival data setting, we make four assumptions:

1 Independent survival times: we expect that the survival time T_i for patient i will not influence the survival time T_j of patient j.

This is a very common assumption that one can find in many statistical applications, including the overall survival field, implying that two records are independent.

2 Non-informative administrative censoring.

The administrative end of follow-up is the date at which we freeze the database and assume all patients whose vital status is not known are censored alive. Non-informative administrative censoring means that patients or tumour characteristics are not predictive of administrative censoring.

An example of informative administrative censoring would be when patients are censored alive at their latest visit to hospital, representing their latest contact with the health system. Patients with better health would have shorter follow-up times. Indeed, healthier patients, who do not need to attend hospital as often as sicker patients, would be censored earlier.

3 Time to death due to cancer, T_E, and due to other causes, T_P, are independent given the knowledge of the demographic variables stratifying the life tables.

In practical terms, we assume that the knowledge of factors such as age, sex, deprivation, ethnicity, region of residence, removes the association between time to death due to cancer and due to other causes. This is true for most cancers. Exceptions exist for specific cancers, such a lung cancer, for which a given patient characteristic, such as smoking, will impact both time to death from cancer and from other causes. [48]

4 Mortality from causes other than cancer is accurately estimated by the life tables, for the population from which cancer patients come from.

Population life tables are the means by which we adjust the overall survival for the impact of causes of death other than cancer. They provide estimates of death rates measured for the population from which cancer patients come from, therefore sharing characteristics such as region of residence, level of deprivation, ethnicity, sex, age and calendar year of death. This assumption holds for as long as cancer patients are not a selected group.

Alternatively, modelling the excess hazard of death gives unbiased estimates of individual excess hazards and individual net survival providing an additional two assumptions are met:

5 The effect of life-table variables are included in the excess hazard model.

The variables stratifying the life tables are generally chosen based on their availability in both the cancer and population data. These variables are key predictors of overall mortality. To adjust for informative censoring, life-table variables need to be included in the excess hazard model. [49]

6 Regression models are correctly specified and contain a meaningful set of variables.

Similar to all forms of regression models, failing to adjust for meaningful predictors of the outcomes, or confounders of the associations between other variables and the outcome will lead to biased estimates of associations and outcomes. Additionally the relationship between each variable and the outcome must be appropriately modelled.

4 Summary

On the one hand, non-parametric estimators of survival are not designed to provide a measure of effect of any predictors in the dataset, but rather estimate net survival by levels of each factor of interest. They are arithmetic calculations using information directly available or added to the cancer registry data such as latest vital status, exact date of death, and expected survival (or background mortality) at time of death.

On the other hand, regression models have been developed to understand the associations between predictors and outcomes. They require further assumptions in order to get a simplified, yet plausible, overview of the reality.

In the context of predictions of survival, we wish to estimate what survival would be for patients outside of the sample of observations available or for whom follow-up has not yet been observed. It may be for other cohorts of patients altogether, or for patients with some characteristics that are not available in the sample. In any case, it would be for unobserved data. The fundamental underlying assumption is that informed predictions of what survival may be for patients outside of the cohort can be made by understanding existing variations in cancer survival. These variations in survival are estimated through the modelling of the excess hazard of death, best achieved using regression models.

In the following chapters, excess hazard models are considered for the estimation of survival for cohorts of cancer patients. We first concentrate on model specification for the estimation of cancer survival (Chapter 1). Next, we explore the available tools necessary to distinguish models in their capacity to predict survival (Chapter 2). Last, we introduce algorithms to choose a set of models that yield good predictions of the patient's cancer survival, for patients outside of the cohort of patients available (Chapter 3).

5 My contributions to the field

I have conducted and been involved in life table methodology in the past few years. We developed methodology to smooth mortality rates derived from sparse data. That methodology enabled us to generate life tables for people of South Asian origin, living in the UK, and by deprivation groups in Portugal. To be useful to cancer survival research, both cancer records and life tables need to be stratified by the same variables.

Construction of South Asian-specific life tables

Maringe C, Li R, Mangtani P, Coleman MP, Rachet B. Cancer survival differences between South Asians and non-South Asians of England in 1986-2004, accounting for age at diagnosis and deprivation. Br J Cancer. 2015;113(1):173-81.

Parametric smoothing of raw estimates of mortality

Rachet B, *Maringe C*, Woods LM, Ellis L, Spika D, Allemani C. Multivariable flexible modelling for estimating complete, smoothed life tables for sub-national populations. BMC Public Health. 2015;15:1240.

Life tables by deprivation in Portugal

Antunes L, Mendonça D, Ribeiro AI, *Maringe C*, Rachet B. Deprivation-specific life tables using multivariable flexible modelling – trends from 2000–2002 to 2010–2012, Portugal. BMC Public Health. 2019;19(1):276.

Correction for mis-matched life tables

Rubio FJ, Rachet B, Giorgi R, *Maringe C*, Belot A. On models for the estimation of the excess mortality hazard in case of insufficiently stratified life tables. Biostatistics. 2019.

Chapter 1

Excess hazard model selection

1.1 Introduction

1.1.1 Statistical models

The primary purpose of a **model** is to represent schematically existing and often complex relationships between different components. A **statistical model** estimates the data-generation process in the population whose observations come from and form a sample of. Assumptions are made to help simplify the structure that specifies the relationships between **predictors** and **outcome(s)**. Predictors are independent variables or inputs, such that no other variables available determine their values. Outcome variables, or outputs, are dependent on the values taken by the predictor variables. Statistical models may be developed with different aims in mind, [50] (1) a **descriptive** purpose: understanding association(s) between factors, (2) an **explanatory** purpose to identify causal relationship(s) or (3) a **predictive** purpose aiming at extrapolating the mechanisms observed in the sample to other samples, outside the data used for estimating the parameters of the model.

Descriptive models aim to reduce the complex reality to a set of associations between predictors and outcome. These models offer a simplified description of existing associations and help understanding of complex mechanisms.

Explanatory models aim to draw causal associations between independent variables and outcomes. One needs a measure of one targeted association of interest, while adjusting for all possible confounders. Explanatory models aim at reducing bias in the estimation of the causal contrast of interest.

Predictive models aim to learn from the existing associations observed in a dataset, to transpose those to another sample of observations, or to the same data but beyond the

period of analysis. The model-based parameters are used to predict the outcome for units of analyses that did not contribute to their estimation. Predictive models aim at reducing an overall prediction error.

The general structure of any type of models is usually based on subject-knowledge or derived from the data. The parameters are estimated from the data. Shape and form of effects of continuous variables can be tested on the data or imposed given prior knowledge.

In the literature, predictions from models can appear under the following terms: extrapolation, out-of-sample projection or forecast. Predictions from a model refer to the estimation of the outcome for covariate patterns that were not necessarily in the training sample. Clinical scores for example refer to predictions although their estimates refer to patients that could have been part of the original sample. These scores give clinicians a tool to communicate their likely outcomes to new patients. In the context of this work, prediction refers to estimating the outcome for values outside the range of values observed in the sample of observations used in model fit. Mainly, we focus on extrapolating calendar and/or follow-up time.

1.1.2 Does a generating model exist?

The 'two cultures' in statistical modelling highlighted by Breiman [51] still represent two forms of modelling conceptions. The *data culture* finds its roots in the idea that a statistical model – one that one could potentially define – is at the origin of any phenomenon. It means that a set of variables and a unique relationship between them explain what is observed. It also flows that one can aim for estimating what this actual model is.

The *algorithmic culture* claims that the data mechanisms are unknown, and a 'true' model does not exist or is too complex to be defined. Within this culture, the goal of statistical modelling is to be able to connect independent variables with outcomes, but without using a (single) statistical model, or any well-defined parametric association. This is where supervised algorithms and machine learning have originated.

In parallel with this distinction between the two cultures, is the distinction between

1. Uniqueness of the generating model, or models that carry similar evidence: are we ready to assume the information available to us is enough to decide on a regression model most likely to have generated the data (data culture)?
2. Accuracy of estimations and simplicity or interpretability: are we most interested in the mechanisms and relationships between variables and outcomes (data culture)?

Or would we be more interested in the values of the predicted outcomes (algorithmic culture)?

3. Data reduction: are we facing an overload of variables in comparison to the number of records available? Do we need to reduce the size of the information or make sense of the main interactions (algorithmic culture)?

Mostly, we remain in the data culture, whereby we rely on well-defined regression model(s) with explicit estimated parameters to summarise the complex web of information between reasonably small sets of predictors and cancer survival. We embrace the challenges posed by the search for an appropriate model: we compare several approaches to model selection and their influence on outcome; we study and derive new tools for qualifying the predictive accuracy of a given model; for prediction, we do not discard regression models of equivalent evidence.

In the algorithmic culture, the functional form of effects and parameters would not be explicit. The estimation of the outcome of interest, such as the contrast or association of interest, is the only output reported. The estimated contrast can be the result of many different algorithms (regression models, classification algorithms, etc.), whose estimates are averaged together. [52]

1.1.3 Modelling cancer survival

Understanding prognosis of cancer patients is key, crucially to patients themselves, but also to clinicians and for public health monitoring and evaluation. Beyond individual patient prognosis, the effects predictors have on cancer mortality may highlight inequalities or unforeseen negative interactions between specific factors, at population level.

Describing survival patterns using non-parametric approaches, as described in Setting, is a first and necessary step. Often these descriptive figures are available through time and help highlight trends and patterns in cancer survival. [53, 54] From these figures, one can draw hypotheses as to what could drive the observed changes in survival. Nonetheless, hypotheses can only be tested through regression modelling since these models estimate the effects of specific predictors on trends and patterns, while adjusting for the potential confounding effects of other factors.

Nonetheless, it is worth mentioning that modelling cancer mortality comes at the cost of additional assumptions that may not always be explicitly stated. Such assumptions can be grouped into the following categories:

1. Distributional: there is a distribution or class of distributions assumed for modelling changes in baseline (excess) hazard of death through follow-up time.
2. Forms of effect: a-priori effects of continuous variables on the outcome may be assumed, such as step, linear or non-linear effects.
3. Effect modifiers: it is possible to allow for interactions between variables.
4. Time-dependence: it is possible to allow for time-dependence of the effect of some variables on the hazard of death.

With the expansion of computer programs to implement ever more complex excess hazard models [6, 46, 55–59] in standard softwares, [7, 8, 60, 61] many applications use flexible parametric models in the literature. [62–69] We witness a shift in presentation from cancer survival figures reported for entire cohorts of patients at given times after diagnosis, to graphs of excess hazard of death in continuous time relative to specific characteristics of individuals. The shift from modelling effects of predictors in follow-up time intervals to modelling these in continuous time allows for less stringent assumptions, and a better adjustment for the effect modification of time. Despite increased complexity in the modelling of the excess hazard of death, not much attention was devoted to best practice in relation to model building. The impact of some features of the models have been reviewed [70–72] but there is no work on the best approach to building an acceptable descriptive excess hazard model and no recommendation on how models may be selected.

We aim to make a contribution in this area, at a time when routine, systematic and structured investigation of possible excess hazard regression models is key. Indeed, population-based datasets on cancer registrations, such as in England, are becoming richer through better completeness of key variables, or routine linkage to other nation-wide datasets from primary and secondary care. The effects of the predictors of cancer mortality – sometimes newly available – need to be appropriately modelled.

We selected two model selection algorithms and we aim to review if and how their use influence the estimation of cancer survival, and lead to better description of the mechanisms underlying cancer mortality.

1.2 Variable and model selection

There are many ways one can choose to screen through possible models. The main groups of techniques for model and variable selection are listed in Table 1.1, with their aim and whether they could be easily implemented in excess hazard modelling. Reviews of these

methods are available. [73, 74] Such reviews tend to highlight modelling difficulties and point to pragmatic solutions; no such recommendations exist yet for excess hazard regression models.

1.2.1 Variable selection: selection of predictors

In many fields where the number of variables examined is very large (such as genetic studies looking at genome associations), the topic of variable selection in model building is extremely relevant. The rational for selecting variables is that a limited number of observations or records can only inform on the effects of a reasonable number of variables on the outcome. Excess hazard regression models rely on the estimation of parameters (i) to characterise the relationship between explanatory variables and outcomes, and (ii) to estimate the baseline excess hazard. There are various event-per-variable recommendations proposed in the literature, based on empirical and simulated work, and depending on the modelling purpose. [75, 76] These range between 10 and 25 events per parameter to be estimated. [77, 78] The topic is also relevant in the competing risks framework where numbers of events (of interest) per parameters estimated influence estimations. [79] Excess hazard models are part of the competing risk framework, which means that among all events observed, only deaths from cancer bring information.

1.2.2 Model selection: selection of the form of effects of predictors

In contrast to variable selection, in which the only concern is whether a variable should be included in a model or not, model selection is about investigations around the most likely functional forms for the effects of continuous factors. In the context of time-to-event data, model selection is also concerned with estimating the baseline hazard of death through time and with identifying which effects are time-varying and how best to model these. The strategies for variable or model selection are the same, except that one tests more complex models including interactions, non-linear and time-dependent effects on the available data.

1.2.3 Introduction to strategies for the selection of relevant effects of predictors

First and foremost, background knowledge of the field as well as a careful and clear definition of the aims of the modelling exercise are necessary before thinking of model selection

Table 1.1: Review of common model selection strategies, and their availability in excess hazard regression models

Type of model selection	Method	Aim	Excess hazard modelling
Variable significance			
	Backward Forward Stepwise	Keep variables that are associated with outcome	Possible
Effect size			
	Change in estimate	Keep confounders of main association	Possible
Shrinkage - penalised likelihood			
	LASSO Elastic net Adaptive LASSO	Penalise effects that are too large	Not available
Kulback-Leibler distance			
	AIC AICc	Minimise the distance between the model and the generating model	Possible
Machine learning			
	Random forest Boosting Bagging	Maximise predictive accuracy	Not available
Resampling			
	Bootstrap Cross-validation	Test for stability	Possible

tools and strategies. [76] Indeed the selection of the best strategy for variable and model selection will depend on the reasons why a regression model is developed in the first place.

Some strategies (Table 1.1) aim at reducing the number of predictors or simplifying the effects of these variables on the outcome in the model (sequential algorithms, information theory, effect size). Others, such as shrinkage methods (penalized regression and the least absolute shrinkage and selection operator (LASSO)), control how large coefficients grow by specifying a shrinkage parameter that control the amount of regularization. That parameter is chosen using cross-validation. [80] To the best of our knowledge, LASSO has not been implemented in excess hazard regression modelling, but regression models using penalised splines (tensor product) have. [5, 56]

Automated approaches, such as iterative model selection, running through potential models in a logical order, such as backward, forward or stepwise, are based on likelihood ratio tests. Provided models are nested, the likelihoods of two consecutive models are compared using a χ^2 distribution with d degrees of freedom. d is the difference in number of parameters between the two models being compared. The model with the largest likelihood is favoured. A small ratio of likelihoods, therefore close to 0, shows strong evidence against the simpler model, meaning it is not likely that it generated the data. Alternatively if the ratio is close to 1, both models are equally likely to have generated the data and there is no evidence against, or to reject, the simpler model. In the context of variable selection, the difference between two models is the presence of a parameter β for the effect of the variable of interest on the outcome. In the context of model selection, the differences between two models can be different parameterisations of a continuous variable. Such models may thus not be nested, and likelihood ratio tests are not valid in such situation.

When models are non-nested, calculating and comparing information criteria between models is possible. Some of these criteria aim to approximate the Kullback-Liebler distance, a measure of closeness (or distance) between two distributions. [81] Laud and Ibrahim [82] proposed a general form of information criteria for the selection of variables:

$$IC(a) = I - a * (k_{m_0} - k_m)$$

I is the likelihood ratio statistics, and k_{m_0} and k_m the number of parameters estimated in models m_0 and m, respectively. Different values of a refer to different criteria such as Box and Kanemasu ($a = 1$), the Bayes factor ($a = 3/2$), Akaike (or AIC, $a = 2$), Schwarz (or BIC, $a = log(N)$), San Martini and Spezzaferri ($a = log(Nc^b)$, $c = 2N\lambda^{-1} * \exp(\frac{\lambda}{N}) - 1$, $b = \frac{2}{k_{m_0} - k_m}$).

If $a \geq 1$, simpler models are favoured over complex models, a desirable property for predictions. The value of some criteria (such as Schwarz, San Martini and Spezzaferri above) are also penalised by the number of subjects N in the analyses. The value of the criteria reflects the distance between two probability distributions or models. It represents how much information is lost when a given model is used over the true (unknown) model. The selected model will be the model that yields the smallest criteria, that is, smallest distance.

The overall aim of model selection is to find a model that is the best possible approximation of the data generation process. When the aim of modelling is to describe, measure and quantify patterns, the chosen models need to be simple and interpretable. Restricted availability of data, coding of variables, sample size, completeness of records, assumptions made, necessity for variable and model selection, will all have an impact on the choice of final model. Many assumptions are made throughout the model building exercise so that it is both attractive, useful and manageable. [83] Although these assumptions are tested on the data, the final estimates that one given model produces carry more uncertainty than it seems or as expressed by the variance estimated around its parameters. Indeed, once the final model is selected, it is generally accepted as the model that generated the data, and the uncertainty related to model selection is not reflected in the final model estimates or inference. [84]

Cross-validation and bootstrap methods can be used for variable selection or act as a validation for the selected model. By testing models on different portions of the data, there is stronger evidence for specific effects of the predictors on outcomes. Cross-validation and bootstrap can also act as sensitivity analyses following any of the methods mentioned above. The *test error* between the predicted model outcomes and the observed outcomes, measured on test samples, is closer to the *true prediction error* than the *training* (or *in-sample* or *apparent*) *error*, measured on the original data. The difference between the *test error* and the *training error* is the *optimism*. The *optimism* reflects how much better a model predicts the outcome on the data it is trained than on an independent sample. A nice feature of the Akaike information criterion (AIC) is that it embeds an estimation of the *test error*, through estimating the *optimism* and adding it to the *training error*. [85]

1.3 Algorithms for functional form selection

We have introduced the concepts of variable and model selection. We insisted that a model aims to remain interpretable and yet be a true reflection of associations between predictors, in the descriptive modelling framework. Mis-specifying the functional form of a predictor may have an impact on the selection of its other functional forms, or of the effects

of other variables' functional forms: this is called self-confounding and confounding. [86] Hierarchical algorithms have been set up, looping through each variable in an iterative way, to ensure the effects selected are correct, independently of the effects of other variables.

We focus on the following two algorithms, developed specifically for time-to-event models, and based on different approaches to model selection, within the spectrum of significance testing: the first algorithm or class of algorithms, developed by Royston and Sauerbrei, is based on forward selection; [87–89] the second algorithm, offered by Wynant and Abrahamowicz, [90] is based on backward selection. Both algorithms are described in details below, before their comparison in the simulation study presented in BMC Methods in Medical Research. [91]

1.3.1 Royston and Sauerbrei algorithm

After describing fractional polynomials, [44] a class of polynomials diverse enough to represent a wide variety of flexible functions, Royston and Sauerbrei developed a series of algorithms specifically for the optimal selection of fractional polynomials in multi-variable model building. [76] Firstly, the close-test procedure concentrates on the selection of the most appropriate degree of freedom for the polynomials. Then, they further extend to restricted cubic splines, [88] to survival models in which time-varying effects are tested, [92] and to interactions between covariates. [87]

These algorithms are neat proposals testing for complex effects of predictor variables, in the presence of other covariates. The initial function selection strategy, the closed-test procedure for function selection, is based on forward selection. The few steps are given below, using restricted cubic splines of a continuous variable x:

1 Choose the most complex splines function of a continuous variable x permitted, say with m knots, $m+1$ degrees of freedom (dfs). $m=0$ is the linear function. The positions of knots are chosen based on subject-matter knowledge or by default (such as centiles of the distribution).

2 The first model fitted is the null model, M_{null}: a model that excludes the continuous variable x.

3 Then, the model with the most complex splines function of x, M_m is fitted.

4 M_m is compared to M_{null}, using a χ^2 test with $m+1$ dfs. The significance level α is fixed a-priori.

5 If the test is non-significant, x has no effect and is excluded.

6 Otherwise, M_m is tested against a model with a linear effect of x, and so on until all models with less knots than M_m are ruled out.

In the presence of many continuous factors, each are investigated in turn, from most to least significant (when evaluated in a model with only linear proportional effects). It means the overall model building algorithm combines as many closed-test procedures as necessary given the number of continuous factors. Binary or categorical factors are also considered for inclusion in the model, and the overall significance of their dummy variables is tested.

With time-to-event data, further models are tested for support from the data, specifically looking at time-varying effects. [92] In the Royston and Sauerbrei framework, time-dependent effects are considered as would interactions be considered, although there is an extra step relative to event-distribution. The tests for non-proportionality are therefore done in two steps, and only once the form of the main effect of each predictor is established. First, the follow-up time is truncated at the mean time to death, and the variable selection procedure described above is performed again on that selected subset of patients. The purpose of this step is to check that no variable with short-term effect only is forgotten. Second, using a stepwise forward approach, each variable in turn is tested for time-varying effect, at the pre-specified significance level α_{TV}, possibly different to α.

Royston and Sauerbrei propose the MFPIgen algorithm that includes testing for interaction, [87, 93] which we also applied on excess hazard models. That algorithm is devised for general interactive effects between two predictors x_1 and x_2, possibly both continuous. The following steps are applied:

1 Apply the algorithm above for the selection of main effects, forcing factors x_1 and x_2 into the model.

2 Calculate the multiplicative effects between the forms selected for x_1 and x_2, leading to k interaction terms.

3 Fit final selected model with the k interaction terms included, and test for these comparing the likelihood ratio to a χ^2 with k degrees of freedom. Perform a joint test of all dummy variables.

4 Check interactions by looking at effects in subgroups, and contrasting with the non-parametric estimates.

5 When more than one interaction is tested, a forward stepwise procedure is used for testing each interaction term in turn.

We adapted this algorithm to the excess hazard model, such that the significance level for testing M_m *vs.* M_0 for life table variables was set to $\alpha = 1$, to take account of informative censoring. We also adapted it to include tests for interactions between categorical (e.g. stage at diagnosis) and continuous (e.g. age and year of diagnosis) variables. [91]

1.3.2 Wynant and Abrahamowicz algorithm

In the time-to-event field, Abrahamowicz and colleagues have been advocating and designing model selection strategies that assess simultaneously non-linear and time-varying effects: the three-step iterative conditional estimation. [86, 90, 94] The hierarchical method for testing these effects simultaneously was tested on a flexible extension of the Estève model for excess hazards. [58, 59] The rational for this strategy is that mis-modelling one of the functional forms impacts the modelling of the other forms of the same effect and of other effects. [86] A formal algorithm, based on an iterative backward elimination procedure, was proposed and compared to simpler, non-iterative approaches in Wynant and Abrahamowicz. [90]

The strategy is based on sequential likelihood ratio tests, to identify the effect that has least support from the data:

1 Fit the most complex model expected, M_C, i.e. including all possible (or permitted, based on subject-matter knowledge) non-linear and time-varying effects. The likelihood of the model is kept in memory.

2 Each effect (non-linearity, non-proportionality, overall) is then removed, one at a time, and the likelihood of the models fitted are kept in memory.

3 Perform likelihood ratio tests for each sub-model in comparison to the original model, M_C.

4 Discard the effect that leads to the highest p-value above the significance level, α. This defines the new most complex model, M_C.

The algorithm stops when all effects yield p-values below α.

We adapted this algorithm to excess hazard modelling, such that the main effect of life table variables are not tested for possible exclusion. We also incorporated interactive effects between some of the predictors, and tested for these in the same way that the algorithm tests for non-linearity or time-varying effects. [91]

In the following published journal article we compare both approaches and provide practical guidance for researchers looking to model cancer (or indeed any disease) survival in the

relative survival data setting for dealing with competing risks. Linear *vs.* non-linear effects of continuous variables, time-dependent effects, and interactions between predictors are investigated.

1.4 Comparison of model-building strategies for excess hazard regression models in the context of cancer epidemiology, Maringe et al., BMC Medical Research Methodology

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	273792	**Title**	Mrs
First Name(s)	Camille		
Surname/Family Name	Maringe		
Thesis Title	On the prediction and projection of cancer survival		
Primary Supervisor	Prof. Bernard Rachet		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	BMC Medical Research Methodology		
When was the work published?	20 November 2019		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I was lead author on the paper. I designed the study, and simulations, in consultation with co-authors, I performed the analyses, interpreted the results and presented the results in the manuscript. I drafted the manuscripts and received comments from all co-authors.

SECTION E

Student Signature	[redacted]
Date	06/02/2020.

Supervisor Signature	[redacted]
Date	5 February 2020

Maringe et al. BMC Medical Research Methodology (2019) 19:210
https://doi.org/10.1186/s12874-019-0830-9

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Comparison of model-building strategies for excess hazard regression models in the context of cancer epidemiology

Camille Maringe[1*], Aurélien Belot[1], Francisco Javier Rubio[2] and Bernard Rachet[1]

Abstract

Background: Large and complex population-based cancer data are becoming broadly available, thanks to purposeful linkage between cancer registry data and health electronic records. Aiming at understanding the explanatory power of factors on cancer survival, the modelling and selection of variables need to be understood and exploited properly for improving model-based estimates of cancer survival.

Method: We assess the performances of well-known model selection strategies developed by Royston and Sauerbrei and Wynant and Abrahamowicz that we adapt to the relative survival data setting and to test for interaction terms.

Results: We apply these to all male patients diagnosed with lung cancer in England in 2012 ($N = 15{,}688$), and followed-up until 31/12/2015. We model the effects of age at diagnosis, tumour stage, deprivation, comorbidity and emergency presentation, as well as interactions between age and all of the above. Given the size of the dataset, all model selection strategies favoured virtually the same model, except for a non-linear effect of age at diagnosis selected by the backward-based selection strategies (versus a linear effect selected otherwise).

Conclusion: The results from extensive simulations evaluating varying model complexity and sample sizes provide guidelines on a model selection strategy in the context of excess hazard modelling.

Keywords: Excess hazard models, Interactions, Non-linearity, Non-proportionality, Variable selection

Background

Population-based cancer datasets have become richer in recent years. Improved completeness of key variables and additional information from linkages with other datasets (secondary care management data, specialised registry data, treatment data) have both contributed to enhance the quality and utility of data. Furthermore, longstanding datasets make possible the analysis of long-term trends and survival probabilities can be estimated further away from the date of diagnosis.

Analysis of population-based cancer survival has greatly benefitted from this data enrichment. However, when modelling the effect of covariates on survival, special care should be taken when assuming or relaxing assumptions of a linear effect or an effect constant in time (the proportional hazards -PH- assumption). Thus, a modelling strategy is required. Aside from the time-to-event setting, many strategies are developed for variable selection and tests for non-linearity of continuous variables, traditionally based on backward, forward or stepwise algorithms. In the time-to-event field in general, and in population-based cancer survival analyses in particular, less attention has been devoted on the selection of the functional form of predictor variables [1, 2]. Indeed, the effects of variables are commonly assumed linear and constant in time, assumptions likely violated for many predictors of cancer survival, especially with long-term follow-up.

Machine learning algorithms have focussed on variables selection in scenarios where tens or thousands of variables are available [3]. These methods mainly focus on factor analysis and random survival forests [4]. In the context of population-based data, the number of

* Correspondence: Camille.Maringe@lshtm.ac.uk
[1]Cancer Survival Group, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK
Full list of author information is available at the end of the article

variables remains low or moderate, but the functional forms of their effects (non-linear and/or time-dependent), as well as their possible interactions need to be carefully examined. Model building fits within three different purposes: descriptive, explanatory and predictive [5]. Our aim here is to describe, measure and quantify accurately the effects of relevant (active) variables while excluding spurious effects.

Some authors [6–8] have shown the importance of taking account as well as testing both non-linearity and time-dependency of effects simultaneously, when modelling time-to-event data, in order to get accurate model-based estimates of survival.

We identify two model-building strategies, developed relatively recently, that offer a systematic and comprehensive approach to the selection of predictors' effects for survival data. One is devised by Sauerbrei and colleagues using fractional polynomials (MFPT) [9] and further adapted for restricted cubic splines (MVRS) [10] and for the inclusion of interactions (MFPI and MFPIgen) [11, 12]. The second one is proposed by Wynant and Abrahamowicz [13], and will be referred to as W&A. These strategies are formulated and tested in the general time-to-event context, in which overall mortality patterns are modelled. Aiming to identify predictors of cancer survival, we focus here on modelling the excess hazard, which is the main quantity of interest in population-based cancer studies [14–16].

Our first aim is to compare and illustrate the use of these model-building strategies (namely MVRS, W&A), in the context of excess hazard regression models. We also propose an extension of those two strategies (called adapted MVRS, aMVRS and adapted W&A, aW&A) for handling interactions between prognostic factors, and compare them to MFPIgen, intended for use with observational data. The performance of these strategies is evaluated in a simulation study mimicking the cancer survival experience of 2000 lung cancer patients diagnosed in 2012 and followed up to the 31/12/2015. We model the effects of explanatory factors on lung cancer survival for the whole cohort of patients diagnosed with lung cancer in 2012. We provide some guidelines over variable and effect selection, based on the simulations.

Methods

a. *The study context: modelling excess mortality*

Our focus is on the excess mortality hazard and the corresponding net survival. The excess mortality hazard is the hazard experienced by cancer patients over and above their background (i.e. expected) mortality hazard due to causes other than the cancer under study. Net survival is derived from the excess hazard and represents the survival experienced by cancer patients under the assumption that they could only die from cancer [17]. Net survival therefore does not depend on the other causes of death, and it is of interest for comparison purposes between countries or periods within a country [18]. In the absence of reliable information on the cause of death, the expected mortality is estimated by the mortality observed in the general population from which patients come from (aka relative survival setting). These life tables are typically defined by age, sex, calendar period, but can also include additional variables such as socio-economic status and ethnicity. Net survival can be estimated non-parametrically [17] or through semiparametric [19] or fully parametric [20–24] excess hazard regression models. Parametric and nonparametric approaches have their own advantages and disadvantages. For the latter, when net survival needs to be estimated in sub-groups, it reduces precision and may lead to unstable estimates. Although there is no assumption relative to the functional forms of effects of variables, these effects cannot be estimated directly. Furthermore, the consistent estimator of net survival proposed by Pohar-Perme and colleagues [17] is unconstrained and thus may show a non-decreasing behaviour in the tails, violating the basic assumptions of survival models. For parametric approaches, the challenges include (a) proper modelling of the baseline excess hazard function, (b) inclusion of potential time-dependent effect of categorical factors, (c) potential non-linear and time-dependent effects of the continuous variables as well as (d) interactions between prognosis factors.

Here, we will use flexible regression models with restricted cubic splines functions for modelling non-linear and time-dependent effects on the log excess hazard scale [23, 25]. The effects of the variables that define the life tables need to be included in the modelling of the excess hazard to produce consistent net survival estimates [17, 20]. Thus, at individual level, the excess mortality hazard $\lambda_E(t,x)$ is linked to the overall $\lambda(t,x)$ and expected (population) mortality hazards $\lambda_P(a+t, y+t, z)$ as follows:

$$\lambda(t,x) = \lambda_E(t,x) + \lambda_P(a+t, y+t, z),$$

where z is a subset of the set of variables x, corresponding to the variables defining the life tables, in addition to age $a+t$ and year $y+t$ (a and y being the age at and year of diagnosis, respectively). The population mortality hazard is considered to be known, and we are interested in estimating $\lambda_E(t,x)$ at time t after diagnosis.

In a general form, the excess hazard regression models considered in our work could be written as follows with

two prognostic variables x_1, continuous, and x_2, categorical (with J categories, $j = 1, \ldots, J$):

$$\lambda_E(t,x) = \lambda_0(t)\,\exp\left(\beta_1(t)*f(x_1) + \sum_{j=2}^{J}\beta_{2,j}(t)*\mathbb{I}_{\{x_2=j\}}\right),$$

where $\lambda_0(t)$ is the baseline excess hazard (defined by using a spline function on the log scale), $f(x_1) = \alpha_1 x_1$ if x_1 is modelled with a linear (L) effect, and f a spline function of x_1 if x_1 is modelled with a non-linear (NL) effect; $\beta_1(t)$ and $\beta_{2,j}(t)$ are spline functions of t if x_1 and x_2 are modelled with time-dependent (TD) effects (the more complicated case), or $\beta_1(t) = \beta_1$ and $\beta_{2,j}(t) = \beta_{2,j}$, $j = 1, \ldots, J$ if not (i.e. PH, the simplest case). For the categorical variable x_2, we considered a "joint" parameterisation of its effect: either all $J - 1$ dummy variables are time-dependent, or none. To simplify notation later, we define $\boldsymbol{\beta_2}(t)*g(x_2) = \sum_{j=2}^{J}\beta_{2,j}(t)*\mathbb{I}_{\{x_2=j\}}$; lastly $\mathbb{I}_{\{x_2=j\}}$ defines an indicator variable (equal to 1 when $x_2 = j$, 0 otherwise).

b. *Model selection strategies*

The MVRS strategy

MVRS is based on an iterative forward selection of variables and increasingly complex functional forms of effects [10]. The model-building proceeds in three steps: (a) the first step focusses on the presence of a variable's effect, and its possible non-linearity in the case of continuous predictors, while assuming proportionality of hazards for all variables. The iterative process loops through all variables from most to least significant, until no effect is removed or added. (b) In the second step, non-proportionality of hazards is explored by restricting the follow-up time to the time until the median time of observed events on which step (a) is performed and additional effects may be retained. (c) The third step consists of testing the non-proportionality of all effects selected in (a) and (b) in a forward stepwise fashion. The likelihood ratio test is used for evaluating significant effects, with a pre-fixed significance level (usually 5%).

The W&A strategy

W&A advocate for the use of an iterative backward elimination of non-significant non-linear and time-dependent effects [13]. From the most complex model, including all possible non-linear and time-dependent effects, each non-linear and time-dependent effect is tested in turn using likelihood ratio test, and the effect corresponding to the highest p-value (above 5%) is removed. From this new model, we test again each remaining non-linear and time-dependent effect in turn, and repeat those steps until all effects kept are significant. The final model is found when all tests yield p-values under 5%.

There are several structural differences in the approaches described above. Firstly, W&A advocates for simultaneous tests of non-linear and time-dependent effects, and the effects are removed one by one, starting from the smallest. By contrast, the MVRS strategy establishes a hierarchy and investigates possible non-linear effects prior to testing time-dependency of the selected effects. The simultaneous tests of effects in W&A may influence subsequent selections of non-linear and/or time-dependent effects. In MVRS, the selection of non-linear effects occurs in the first step, which may well influence the later selection of time-dependent effects, but the selection of time-dependent effects will not affect retention of non-linear effects. Secondly, the initial models considered are different and lead to backward (in the case of W&A) or forward (MVRS) selection of variables.

Strategies in the relative survival setting

In both strategies, the main life table variables (age, sex, year and deprivation) are forced into the models, as recommended for excess hazard regression modelling [14, 17, 20]. For the non-life table variables linearity and time-dependency and overall effects are tested so the variables could be completely removed from the set of predictors.

Extensions of the strategies for testing for interactions

The authors of MVRS also consider interactions between variables retained, once the main effects have been selected [11]. MFPI and MFPIgen are defined to consider categorical-by-continuous interactions and continuous-by-continuous interactions respectively, even though (from our understanding) they do not test for non-proportionality of the interaction terms [9].

We propose to adapt the original W&A and MVRS strategies to include tests for the form and presence of interactions in the same fashion that they already test for the functional form and inclusion of each variable.

There are three types of possible interactions: between two continuous variables, between a continuous and a categorical variable, and between two categorical variables. We focus on continuous–by-categorical interaction, and the strategies will need to test whether or not the interaction is needed and if it is time-dependent.

The general form of the excess hazard model is as follows, with x_1 continuous and x_2 categorical (with J categories $j = 1, \ldots, J$):

$$\lambda_E(t,x_1,x_2) = \lambda_0(t)\,\exp\big(\beta_1(t)*f(x_1) + \boldsymbol{\beta_2}(t)*g(x_2) + \boldsymbol{\beta_3}(t)*f(x_1)*g(x_2)\big),$$

with all functions as defined above.

The adapted version of MVRS, aMVRS, tests for each interaction in the three steps presented earlier: (a) joint test of the interaction factors, i.e. test for $\boldsymbol{\beta_3} = 0$; (b) In the restricted follow-up time (until the median time of observed events) significance test for $\boldsymbol{\beta_3} = 0$; (c) If $\boldsymbol{\beta_3} \neq 0$ in either (a) or (b), test time-dependence of the interaction, i.e. $\boldsymbol{\beta_3}(t) = \boldsymbol{\beta_3}$.

The adapted version of the W&A algorithm, aW&A, tests for each interaction in the same way it tests for the effects of main variables: it first tests for time-dependent effect of the interaction, i.e. $\boldsymbol{\beta_3}(t) = \boldsymbol{\beta_3}$, and then, if a time-fixed effect is favoured, it tests for the main effect of the interaction $\boldsymbol{\beta_3} = 0$.

The MFPIgen algorithm only considers interactions in a final step, after selecting the main effects of variables in the usual steps (a)-(c). It tests for $\boldsymbol{\beta_3} = 0$. In all algorithms the forms of the interactions, f and g are defined by the form of the main variables x_1 and x_2 as they are modelled when the interaction is considered.

In the case of interactions with categorical variables, the presence of the interaction could be tested in two different ways: overall (called joint test [26]), or each level of the interaction separately. Here we only test the interactions as one effect, such that all factors relating to one interaction would be removed/included when testing for their inclusion. In the algorithms, the user specifies which interaction terms are worth investigating. Specific significance levels for the tests related to interactions may be chosen as in MVRS. Additional file 1 details how the algorithms are adapted to testing for interactions.

c. *Simulation of biologically plausible lung cancer survival data*

Data generation and simulations design

We use the observed survival time and vital status of the full cohort of lung cancer patients (N = 17,597), evaluated on the 31st December 2015, to obtain the regression coefficients of an excess hazard regression model. The large sample size enables detection and precise estimation of small effects. These coefficients are used for simulating cancer survival times, as detailed in formulas (A)-(D) below. From this excess hazard regression model, the cancer survival time Tc is generated using the inverse transform method [27, 28].

For the data design, we randomly extract 2000 men diagnosed with lung cancer in England in 2012 from the English population-based cancer registry, among those with valid information on stage at diagnosis. We kept the information on their age at diagnosis (continuous variable), their level of deprivation (categorical variable with 5 levels of increasing deprivation measured by the income domain of the Index of Multiple Deprivation [29]), and their stage of cancer at diagnosis (categorical variable with 4 levels of increasing severity based on the Tumour, Nodes, Metastasis classification [30]). The relatively small sample size for population-based data will enable us to test the practical performances of the algorithm in a setting with low censoring rate (15%) but small number of patients (relative to standard population studies). We repeated this for a larger sample of 5000 cancer patients to study the sensitivity of the model selection strategies on the number of events. By default, all results are presented for the samples of 2000 patients, except when clearly mentioned.

We devise four simulation scenarios, representing increasingly complex excess hazard regression models (see Box 1):

(A) Model with linear and proportional effect of age, and proportional effects of stage and deprivation, without interaction

$$\lambda_E(t, age, stage, dep) = \lambda_0(\ln(t)) * \exp\left(\alpha * age + \sum_{i=2:4} \beta_i * \mathbb{I}_{stage=i} + \sum_{j=2:5} \gamma_j * \mathbb{I}_{dep=j}\right).$$

(B) Model with linear and proportional effect of age, and proportional effect of stage, deprivation and an interaction between age and stage

$$\lambda_E(t, age, stage, dep) = \lambda_0(\ln(t)) * \exp\left(\alpha * age + \sum_{i=2:4} \beta_i * \mathbb{I}_{stage=i} + \sum_{j=2:5} \gamma_j * \mathbb{I}_{dep=j} + \sum_{k=2:4} \alpha_k * age * \mathbb{I}_{stage=k}\right).$$

(C) Model with non-linear and time-dependent effects of age, time-dependent effects of stage, and proportional effects of deprivation, without interaction

$$\lambda_E(t, age, stage, dep) = \lambda_0(\ln(t)) * \exp\left((\alpha + \alpha^* * \ln(t)) * f(age) + \sum_{i=2:4} (\beta_i + \beta_i^* * \ln(t)) * \mathbb{I}_{stage=i} + \sum_{j=2:5} \gamma_j * \mathbb{I}_{dep=j}\right).$$

(D) Model with non-linear and time-dependent effects of age, time-dependent effects of stage, proportional effects of deprivation and a proportional interaction between age and stage

$$\lambda_E(t, age, stage, dep) = \lambda_0(\ln(t)) * \exp\left((\alpha + \alpha^* * \ln(t)) * f(age) + \sum_{i=2:4} (\beta_i + \beta_i^* * \ln(t)) * \mathbb{I}_{stage=i} + \sum_{j=2:5} \gamma_j * \mathbb{I}_{dep=j} + \sum_{k=2:4} f(age) * \mathbb{I}_{stage=k}\right).$$

In the formulas above, associated to scenarios A-D, f denotes a restricted cubic splines function with 2 degrees of freedom, i.e. 1 internal knot placed at the

Box 1 Summary of the effects simulated

	Age	Stage	Deprivation	Age*Stage
A	L-PH	PH	PH	–
B	L-PH	PH	PH	PH
C	NL-TD	TD	PH	–
D	NL-TD	TD	PH	NL-PH

L Linear, *NL* Non-linear, *TD* Time dependent, *PH* Proportional hazards

median age of the patients' cohort, $\lambda_0(ln(t))$ is a restricted cubic spline function of time, with up to 3 degrees of freedom, i.e. 2 internal knots placed at the tertiles of the distribution of times to death.

Time to death from other causes Tp is generated assuming a piecewise exponential hazard obtained from general population life tables detailed by month of age, sex, calendar month and deprivation level [20]. The censoring time C is evaluated on 31/12/2015. The final observed follow-up time for each individual is defined as $T = min\ (Tc, Tp, C)$, with the corresponding vital status indicator δ (i.e., $\delta = 0$ for censored observations and $\delta = 1$ for death).

For each scenario (A-D), we simulate 250 datasets, and we utilise the `survsim` command in Stata [28] for simulating cancer survival times in the scenarios described above. Close to 90% of patients die in the first four years after diagnosis, classifying lung cancer in the poor-prognosis cancers with low censoring rate.

Analysis of simulated data

The classical algorithms, W&A and MVRS, are run on scenarios A and C, while the algorithms extended to testing for interactions, aW&A and aMVRS, are run on scenarios B and D. MFPIgen is also tested on scenarios B and D. All excess hazard regression models are fitted using the `strcs` command in Stata [25], as described in section 2.a.

d. *Indicators used for comparing the model-building strategies*

One additional binary variable not contained in the life tables and absent from the original simulation models is added when testing the model-building strategies. For each scenario, we compare the models selected by each strategy to the original effects used in the simulation with the following indicators.

Firstly, we summarise the proportions of models that select each variable with their non-linear or time-dependent effects for each algorithm. We also study the confounding and self-confounding effects: the impact of mis-specifying one of the components (TD, NL, interactions) of the functional form of a covariable on its other components or on the selection of such components for other variables. We also calculate the proportion of selected models that contain or are exactly the simulated models for each strategy.

Furthermore we provide sensitivity (true positive) and specificity (true negative) values, as defined below, looking at the number of correctly selected effects and the number of correctly unselected effects over the number of active and inactive effects [31]. Both sensitivity and specificity tend to reach 1 for a good estimator:

$$Se = \frac{\#correctly\ selected\ effects}{\#active\ effects}$$

$$Sp = \frac{\#correctly\ unselected\ effects}{\#inactive\ effects}$$

Then, for each model building strategy we plot the average of the 250 stage-specific cohort net survival curves and compare them to the true net survival curve. We quantify this comparison by calculating the proportion of the Area Between Curves through time, *pABCtime* [32]. *pABCtime* represents the area between each individual net survival curve (or the average of the 250 net survival curves) and the true generating net survival curve (the reference function). It is expressed as a proportion of the area under the true net survival curve (area under the reference function). A *pABCtime* of 0 % means that the cohort net survival estimates under investigation are in perfect agreement with the true initial observed effect.

For any function f, let us assume that the true generating function f^* and the estimated function $\hat{f}$ cross at time t^*, *ABCtime* is defined as

$$ABCtime = \left|\int_0^{t^*} f^*(u)du - \int_0^{t^*} \hat{f}(u)du\right| + \left|\int_{t^*}^{T} f^*(u)du - \int_{t^*}^{T} \hat{f}(u)du\right|,$$

and *pABCtime* as

$$pABCtime = \frac{\left|\int_0^{t^*} f^*(u)du - \int_0^{t^*} \hat{f}(u)du\right| + \left|\int_{t^*}^{T} f^*(u)du - \int_{t^*}^{T} \hat{f}(u)du\right|}{\int_0^T f^*(u)du}.$$

pABCtime is also calculated for the excess hazard curves estimated for given patients' factors and for the effects of age, deprivation, and stage comparing the possibly time-dependent estimated HR curves to the originally simulated HR. In such instances, $\hat{f}$ represent the excess hazard, $\hat{f}(u) = \lambda_E(u, age, stage, dep)$ or excess hazard ratio, $\hat{f}(u) = \exp(\hat{\beta}(u))$.

We also provide bias of effects, at specific time points t_k, which are the average bias over all samples (M = 250) between the estimated (possibly time-dependent) effects of age, stage and deprivation and their simulated effects.

We specify monthly t_k, from diagnosis through to the end of follow-up (4 years):

$$bias\left(\widehat{\beta(t_k)}\right) = \frac{1}{M}\sum_{k=1}^{M}\left(\beta(t_k) - \widehat{\beta(t_k)}\right).$$

e. *Application*

We apply the five model-selection strategies (MVRS, W&A, MFPIgen, aMVRS, and aW&A) to our full cohort of 15,688 men diagnosed with non-small cell lung cancer in 2012 in England and followed-up until 31/12/2015. All patients had a minimum potential follow-up of 3 years. Patient's information on age, deprivation, survival time and vital status is enhanced by information on stage at diagnosis [33] coded using the TNM system (I-IV), emergency route to diagnosis (binary variable) [34], comorbidity status defined after ascertainment of hospital episodes in the 6 months to 6 years prior to diagnosis (binary variable) [35]. The model building strategies test the main effects as well as interactions between age at diagnosis and all other covariates.

All model building strategies yield very similar models (Table 1): no main effect is removed, time-dependent effects of stage, comorbidity and emergency presentation are kept, and when tested, interactions between age and comorbidity is removed by the MVRS algorithm and age and comorbidity and age and emergency presentation by the aW&A and MFPIgen algorithms. Non-linear time-dependent effects of age are retained by the W&A and aW&A algorithms in comparison to linear time dependent effects of age retained in all other model selection algorithms.

Figure 1 illustrates the impact the different selected interactions and linearity/non-linearity of age have on the estimated net survival probabilities for two patients, aged 60 and 80 respectively with the values of other variables set (i.e. stage III, non-emergency presentation, no comorbidity, least deprived). The curves for W&A and MVRS overlap. The selection of interactions in the model impacts the estimated individual excess hazard and cancer survival: there are smaller differences in excess hazard between patients aged 60 and 80 when no interactions are modelled, compared to when interactions are considered. We super-imposed the non-parametric estimator of net survival (red curves) estimated for the 165 patients aged]50–70[years (mean age 64) and the 130 patients aged]75–85[years at diagnosis (mean age 79), with non-emergency presentation, stage III disease and from the least deprived group of the population. The non-parametric net survival estimates are generally lower than all model-based estimates from 1 year (age 80) and 2.5 years (age 60) after diagnosis. At the start of follow-up, the non-parametric estimates tend to resemble the model-based estimates without interaction terms.

These differences at individual level do not however impact the overall cohort estimate of net survival as shown by the hardly distinguishable curves in Fig. 2, similar to the non-parametric estimator of net survival.

Results

The four simulated scenarios represent increasingly complex but realistic excess hazard models, derived from observed records of lung cancer patients. To assess how realistic these scenarios are, we compare the model-based cohort estimates of net survival (using the model used for each simulated scenario) to the non-parametric Pohar-Perme estimates (Additional file 2) on the original, observed data. All scenarios show reasonable stage-specific cohort net survival estimates. Scenarios A and B under-estimate net survival until 12–24 months for patients diagnosed at stages I-III because of the

Table 1 Statistically significant effects of selected prognostic factors identified with each of the five alternative model-building strategies

Variables	aMVRS	MVRS[a]	aW&A	W&A[a]	MFPIgen[b]
Age	L-TD	L-TD	NL-TD	NL-TD	L-TD
Stage	TD	TD	TD	TD	TD
Deprivation	PH	PH	PH	PH	PH
Comorbidity	TD	TD	TD	TD	TD
Emergency diagnosis	TD	TD	TD	TD	TD
Age*Stage	PH	–	PH	–	PH
Age*Deprivation	PH	–	PH	–	PH
Age*Comorbidity		–		–	
Age*Emergency diagnosis	PH	–		–	

L Linear, *NL* Non-linear, *TD* Time dependent, *PH* Proportional hazard
[a]Interactive effects not tested
[b]Main effects from MVRS strategy before testing for interaction

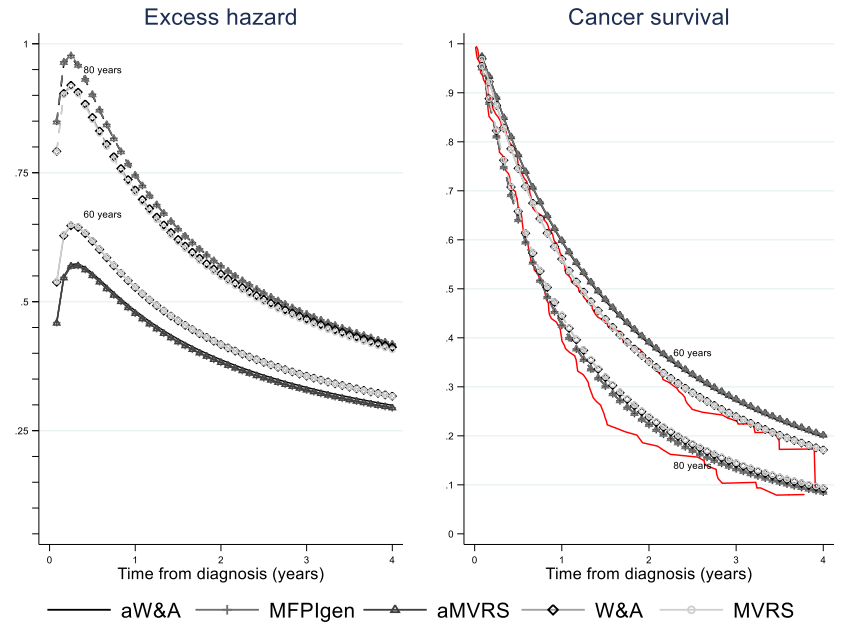


Fig. 1 Excess hazard and survival curves estimated for two patients[1] aged 60 and 80 years at diagnosis: impact of model selection.
[1]patient with the following characteristics: stage III, non-emergency presentation, no comorbidity, least deprived. Plain red curves show the non-parametric estimator of net survival for patient aged 50–70 (upper curve) or 75–85 (lower curve) years at diagnosis

simple effects modelled. Scenarios C and D include non-proportional effects of the main factors and estimate stage-specific cohort net survival very neatly. The characteristics of the patients used in the simulations are presented in Additional file 3. Patients in stage IV comprise half of the sample. There is a decreasing average age with increasing stage at diagnosis. The distribution of patients by deprivation group is skewed towards more deprived groups, and a third of the patients have the trait of the extra binary variable.

(a) *Performances of the model-building strategies in selecting variables and their effects*

Original algorithms – scenarios A and C (no interaction)
In scenario (A), both algorithms led to almost identical selection of effects (Fig. 3, Table 2). The only difference is the higher proportion of time-dependent effects of the extra variable, 5.6% vs. 0.8%, selected with W&A compared to MVRS. In scenario (C), albeit small there are more differences in the effects selected between

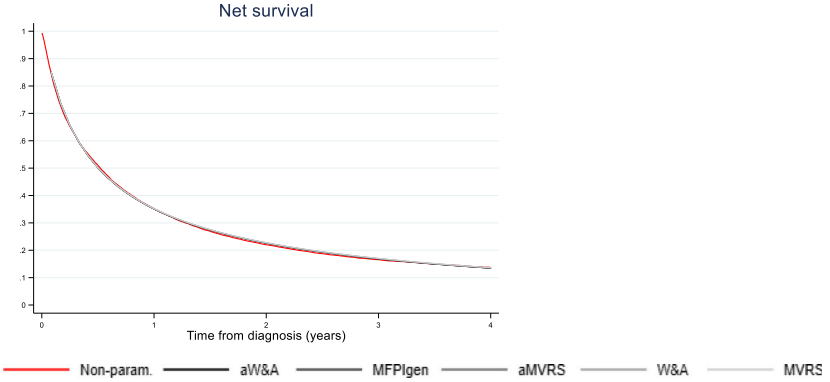


Fig. 2 Net survival for the cohort of 15,688 men with lung cancer

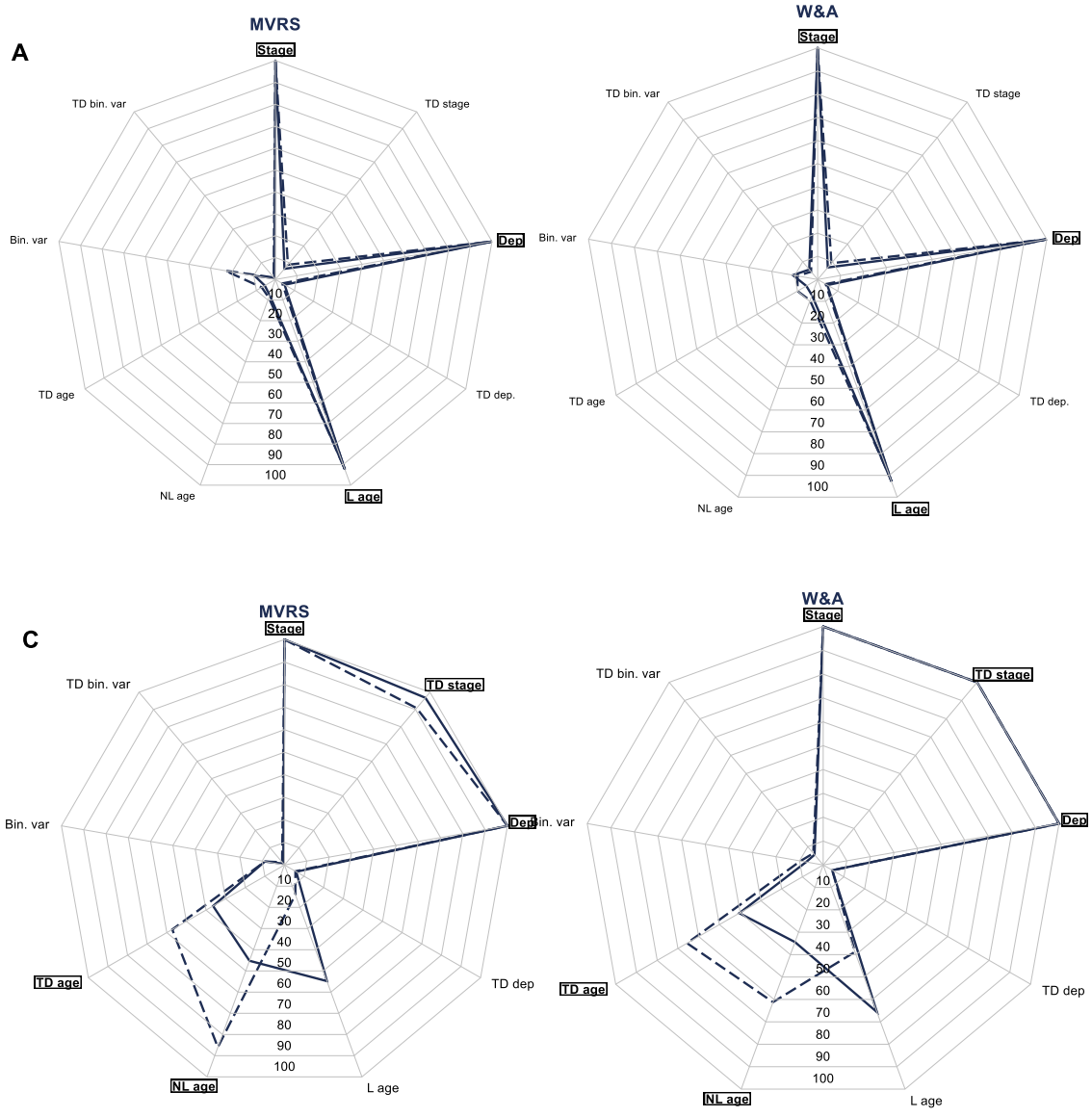


Fig. 3 Variables and effects (linear/non-linear, proportional/time-dependent) selected: scenarios A and C, samples of 2000 (plain line) and 5000 (dashed line) patients

MVRS and W&A. W&A tends to (rightly) select more models that include time-dependent effects of stage (100% vs. 96.8%) and age (40.4% vs. 36.4%). Non-linear effects of age are more often selected by MVRS (45.2%) than by W&A (34.4%). Overall, the effect of stage is always rightly kept in the final selected models, by all algorithms, and the extra binary variable appears (wrongly) in only 7.2–8.8% of models (Fig. 3, Additional file 4).

All selected models contain the true simulated model for scenario A but the proportions drop to 69.6% (MVRS) and 70.4% (W&A) of models that are the exact simulated model. Similarly in the slightly more complex scenario (C), 10.8% of models contain, and 8.8% of models are, the true model using MVRS model selection, vs. 6.0 and 5.2% of W&A models, respectively (Table 2). This drop in proportions between scenarios A and C reflects the high proportion of models with a time

Table 2 Summary of models and variables selected by each algorithm, on 250 samples of $N = 2000$ and $N = 5000$ patients: scenarios A-D

	Overall model									Sensitivity			Specificity		
	Contained			Correctly selected			Almost correctly selected*			mean	min	max	mean	min	max
A	p (%)	95% CI**		p (%)	95% CI**		p (%)	95% CI**							
***N* = 2000**															
MVRS	100.0	*100*	*100*	69.6	*64.0*	*75.2*				0.97	0.67	1.00	0.89	0.75	0.92
W&A	100.0	*100*	*100*	70.4	*64.8*	*76.0*				0.98	0.67	1.00	0.89	0.67	0.92
C															
MVRS	10.8	*7.0*	*14.6*	8.8	*5.3*	*12.3*	82.8	*78.2*	*87.4*	0.74	0.60	0.80	0.88	0.70	0.90
W&A	6.0	*3.1*	*8.9*	5.2	*2.5*	*7.9*	91.6	*88.2*	*95.0*	0.74	0.60	0.80	0.88	0.60	0.90
B															
aMVRS	35.6	*29.8*	*41.5*	14.4	*10.1*	*18.7*	53.6	*47.5*	*59.7*	0.80	0.50	1.00	0.85	0.45	0.91
MFPIgen	29.6	*24.0*	*35.2*	14.4	*10.1*	*18.7*	66.8	*61.1*	*72.6*	0.81	0.50	1.00	0.88	0.64	0.91
aW&A	35.2	*29.4*	*41.0*	14.8	*10.5*	*19.1*	45.2	*39.1*	*51.3*	0.79	0.50	1.00	0.84	0.36	0.91
D															
aMVRS	2.4	*0.5*	*4.3*	1.6	*0.1*	*3.1*	16.0	*11.5*	*20.5*	0.56	0.50	0.83	0.80	0.33	0.89
MFPIgen	3.2	*1.1*	*5.4*	2.8	*0.8*	*4.8*	36.8	*30.9*	*42.7*	0.59	0.50	0.83	0.85	0.56	0.89
aW&A	4.8	*2.2*	*7.4*	1.6	*0.1*	*3.1*	23.2	*18.1*	*28.4*	0.57	0.50	0.83	0.75	0.22	0.89
***N* = 5000**															
	Overall model									Sensitivity			Specificity		
	Contained			Correctly selected			Almost correctly selected*			mean	min	max	mean	min	max
A	p (%)	95% CI**		p (%)	95% CI**		p (%)	95% CI**							
MVRS	100.0	*100*	*100*	56.4	*50.4*	*62.5*				0.97	0.67	1.00	0.88	0.67	0.92
W&A	100.0	*100*	*100*	68.0	*62.3*	*73.7*				0.97	0.67	1.00	0.89	0.67	0.92
C															
MVRS	46.0	*39.9*	*52.1*	40.8	*34.8*	*46.8*	78.8	*73.8*	*83.8*	0.78	0.60	0.80	0.88	0.60	0.90
W&A	29.2	*23.7*	*34.8*	26.0	*20.7*	*31.4*	87.6	*83.6*	*91.6*	0.80	0.60	0.80	0.88	0.60	0.90
B															
aMVRS	67.9	*62.2*	*73.6*	37.3	*31.5*	*43.3*	55.8	*49.7*	*61.9*	0.80	0.50	1.00	0.87	0.64	0.91
MFPIgen	28.0	*22.5*	*33.5*	14.4	*10.1*	*18.7*	65.6	*59.8*	*71.4*	0.87	0.50	1.00	0.85	0.36	0.91
aW&A	69.2	*63.6*	*74.8*	37.6	*31.7*	*43.5*	55.6	*49.5*	*61.7*	0.86	0.50	1.00	0.82	0.09	0.91
D															
aMVRS	34.4	*28.6*	*40.2*	13.6	*9.4*	*17.8*	24.8	*19.5*	*30.1*	0.66	0.33	0.83	0.79	0.33	0.89
MFPIgen	28.0	*22.5*	*33.5*	18.4	*13.7*	*23.1*	35.6	*29.8*	*41.5*	0.69	0.50	0.83	0.84	0.56	0.89
aW&A	22.0	*17.0*	*27.1*	13.6	*9.4*	*17.8*	34.8	*29.0*	*40.6*	0.65	0.50	0.83	0.73	0.00	0.89

* model C: relaxed NL and TD of age; B: relaxed interaction age*stage; D: relaxed NL and TD of age

** formula for the 95% confidence intervals, with $z = 1.96$ and $w = 250$: $\frac{\hat{p}+\frac{z^2}{2w}}{1+\frac{z^2}{w}} \pm \frac{z}{1+\frac{z^2}{w}}\sqrt{\frac{\hat{p}(1-\hat{p})}{w}+\frac{z^2}{4w^2}}$, using the Wilson approximation [36]

dependent effect of linear age, in other words the low proportion of models with a time-dependent effect of non-linear age. This is explained by the small sample size and the relatively small magnitude of the non-linear and time-dependent effects of age (Additional file 5). Higher number of lung cancer patients leads to higher proportions of selected models that contain or are exactly the generated model (Table 2) due to higher proportions of models capturing the non-linearity and time-dependency of age (Fig. 3).

Sensitivity and specificity are high for scenario A, and are not impacted by an increasing sample size. They are relatively high for scenario C, with a slight increase in sensitivity (0.74 to 0.78–0.80) with an increasing sample size (Table 2).

Algorithms adapted to models with interactions – scenarios B and D

The adapted (aMVRS, aW&A) and MFPIgen algorithms correctly keep the main effects in the final models

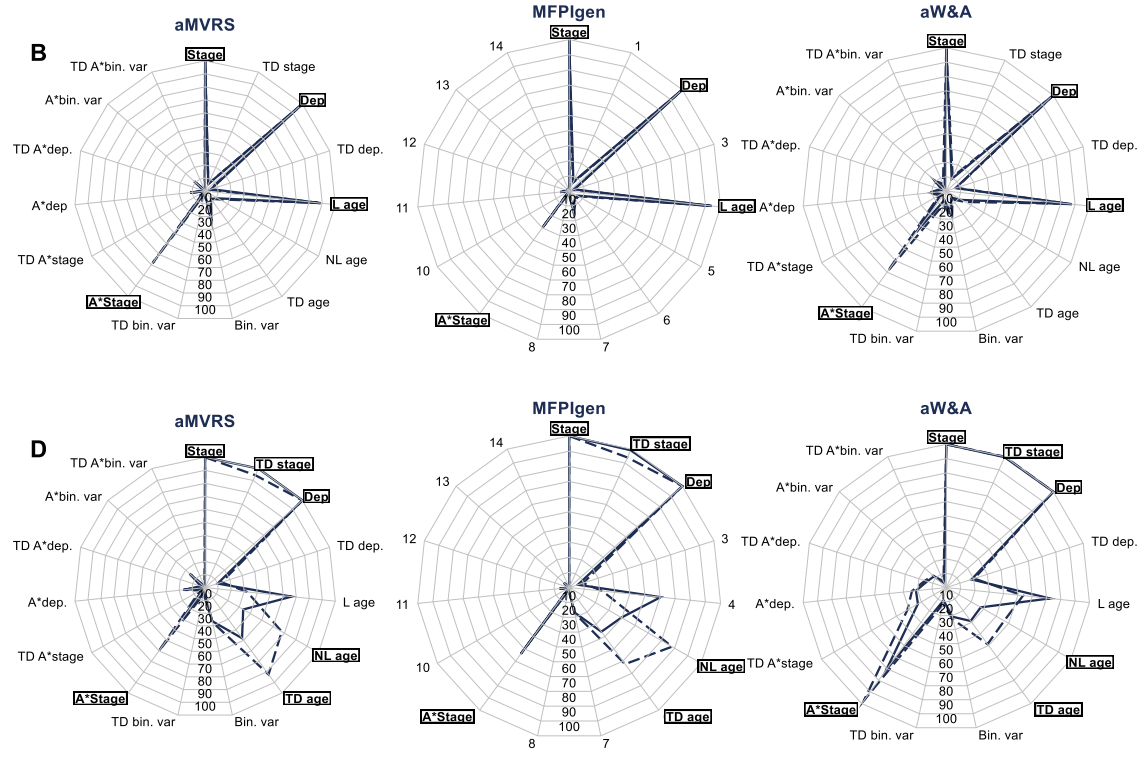


Fig. 4 Variables and effects (linear/non-linear, proportional/time-dependent) selected: scenarios B and D, samples of 2000 (plain line) and 5000 (dashed line) patients

(Fig. 4). 28% of models selected using aW&A identify the non-linearity of age in D, whereas 34–40% of the aMVRS and MFPIgen algorithms retain the non-linearity of age. The aW&A algorithm tends to keep higher proportions of time-dependent effects of deprivation, of the binary variable and of interactions than the other two algorithms. aMVRS and aW&A also lead to 10–21% of interactions wrongly selected. The proportions of the interaction age-stage rightly kept are at or just over 30% for scenario B and up to 71% (aW&A) for scenario D. The MFPIgen algorithm is able to keep in valid interaction in 29.6% (B) and 50.8% (D) of the final models while spurious interactions are rejected in over 94% of final models.

Non-linearity and time-dependency of age in scenario D are retained in just over a quarter of models selected by aW&A, 6–20% less than the proportions of models selected by aMVRS and MFPIgen that contain these characteristics of age. Increased sample size to $N = 5000$ is beneficial for raising the detection of the age-stage interaction in B for aMVRS (68.3%) and aW&A (69.2%), and raising detection of non-linearity and time-dependency of age in D for all three algorithms (Fig. 4).

The proportions of models that contain the true generating model lie between 29.6% (MFPIgen) and just over 35% (aMVRS and aW&A) for scenario B, and between 2.4% (aMVRS) and 4.8% (aW&A) for scenario D. For scenario B, those proportions correspond to the proportion of models with an age by stage interaction, and therefore increase with increasing sample size for aMVRS (74.5% for B and 43.5% for D when N = 5000) and aW&A (72.3% for B and 17.4% for D when N = 5000). For scenario D, this is the proportion of models with an interaction between a non-linear effect of age and stage. Only 14.4–14.8% (scenario B) and 1.6–2.8% (scenario D) are the exact simulated models. These proportions increase to 16–36.8% (scenario D) when small effects are not considered, due to the relatively small sample size, or when the sample size is increased to 5000.

Sensitivity and specificity are around and over 0.8 for scenario B and are stable to increased sample size.

Sensitivity is just over 0.5, and specificity between 0.75 and 0.85 for scenario D, with slight improvement in sensitivity with increased sample size (Table 2).

Impact of mis-selection of effects on other effects

In scenario (A) and (C), W&A seems to suffer more from confounding and self-confounding (Additional file 4). For example, when the extra binary variable is selected in (C), the proportion of models with time-dependent effects of deprivation and/or age are hardly changed with MVRS, but they increase with W&A to 16.7% (+ 12.3%) and 55.6% (+ 15.2%) respectively. (Additional file 4).

There are hardly any confounding or self-confounding effects in the MFPIgen algorithm. Mis-specification of time-dependent effects only has minimal confounding impact on the other effects selected using the aMVRS algorithm. This is due to the two-step structure of the algorithm (Additional file 4).

In the aW&A algorithm, selection of complex forms (e.g. time-dependent effect of a variable) results on the selection of more complex effects of some other factors or additional selection of interaction terms (Additional file 4).

(b) *Accuracy of the non-linear and time-dependent effects estimated*

Original algorithms – scenarios A and C (no interaction)

Figure 5 presents the effects estimated by the models selected following the MVRS or W&A algorithms together with their averaged effects (black line) compared to the true generating effect (red line). All sample sizes are $N = 2000$ patients.

Although there are varied sizes of effect estimated as shown by the width of the boxes (effects estimated as fixed in time) and the varied shapes of the individual effects, grey curves (time-dependent effects estimated), the average effects generally agree with the generating effects for all strategies, and lead to comparable estimated effects for MVRS and W&A. For both strategies, the effects of age are well captured for scenario (A) and (C): *pABCtime* values are 0.3% (A), 0.3% (C, MVRS) and 0.2% (C, W&A), Table 3.

The mixture of time-fixed and time-dependent effects of stage estimated in the selected models for scenario (A) leads to a very good estimation of the average effect compared to the generated effect for both strategies. Note the graphs present log hazard ratios for better illustrating the differences, but *pABCtime* values are calculated on the areas between the hazard ratio curves. *pABCtime* values for the hazard ratios are very similar between algorithms, highest for stage IV (2.5%), intermediate for stage II (2.2–2.4%) and lowest for stage III (1.7%). In scenario (C) all estimated effects are time-dependent, and most shapes agree with the original effect. *pABCtime* values are slightly lower for the W&A algorithm compared to MVRS: 2.3% vs. 2.4% at stage II vs. I, 0.9% vs. 1.2% at stage III vs. I, and 1.8% vs. 2.1% at stage IV vs. I.

The effects of deprivation are well estimated by all models selected by all algorithms: *pABCtime* is below 1.2% for all deprivation categories, and in both scenarios A and C.

More complex effects of the extra binary variables are captured by W&A, in both (A) and (C) leading to slightly higher *pABCtime* values: 0.6% vs. 0.3% (A) and 0.12% vs. 0.08% (C).

Algorithms adapted to models with interactions – scenarios B and D

Figure 6 displays the effects estimated by the selected models (250 grey curves) following the aMVRS, MFPIgen and aW&A algorithms together with their averaged effects (black line) compared to the true generating effect (red line). The effects of age are now split by stage at diagnosis, since an interaction age-stage is simulated.

For all selected models, the average HRs for age seem to generally underestimate the simulated effects for stages I-II, in scenario B and D. These are reflected by larger stage-specific *pABCtime* values for age: 2.4–5.9% (stages I-II) versus 0.01%-2.2 (stages III-IV, Table 3). The time-dependency of age, simulated in scenario D, is not very strong, hence the many models that selected a time-fixed effect for age. Graphs of the non-linear effects of age at given times after diagnosis are presented in Additional file 5.

The effects of stage, deprivation (Fig. 6) and the additional binary variable (Additional file 6) are well reproduced by the average effects obtained from the selected models. The *pABCtime* values can hardly distinguish between the performance of the model-selection algorithms (Table 3). The complexity of models selected by the aW&A algorithm does not impact the overall measures of effects and their adequacy to describe the true generating effects. Indeed, none of the modelled time-dependent effects are strong, but the results presented here shed some light in terms of the sensitivity of the different model selection tools.

(c) *Estimation of the cohort net survival*

For all model-building strategies, the estimated stage-specific cohort net survival curves lie around the original estimated cohort net survival curves, for all subgroups defined by stage at diagnosis for scenarios A-D (Fig. 1). All *pABCtime* values are below 1.7% (Table 3).

The outcome of choice – net survival – is well reproduced by models selected by each strategy and provides reassurance that the experience of cancer survival for

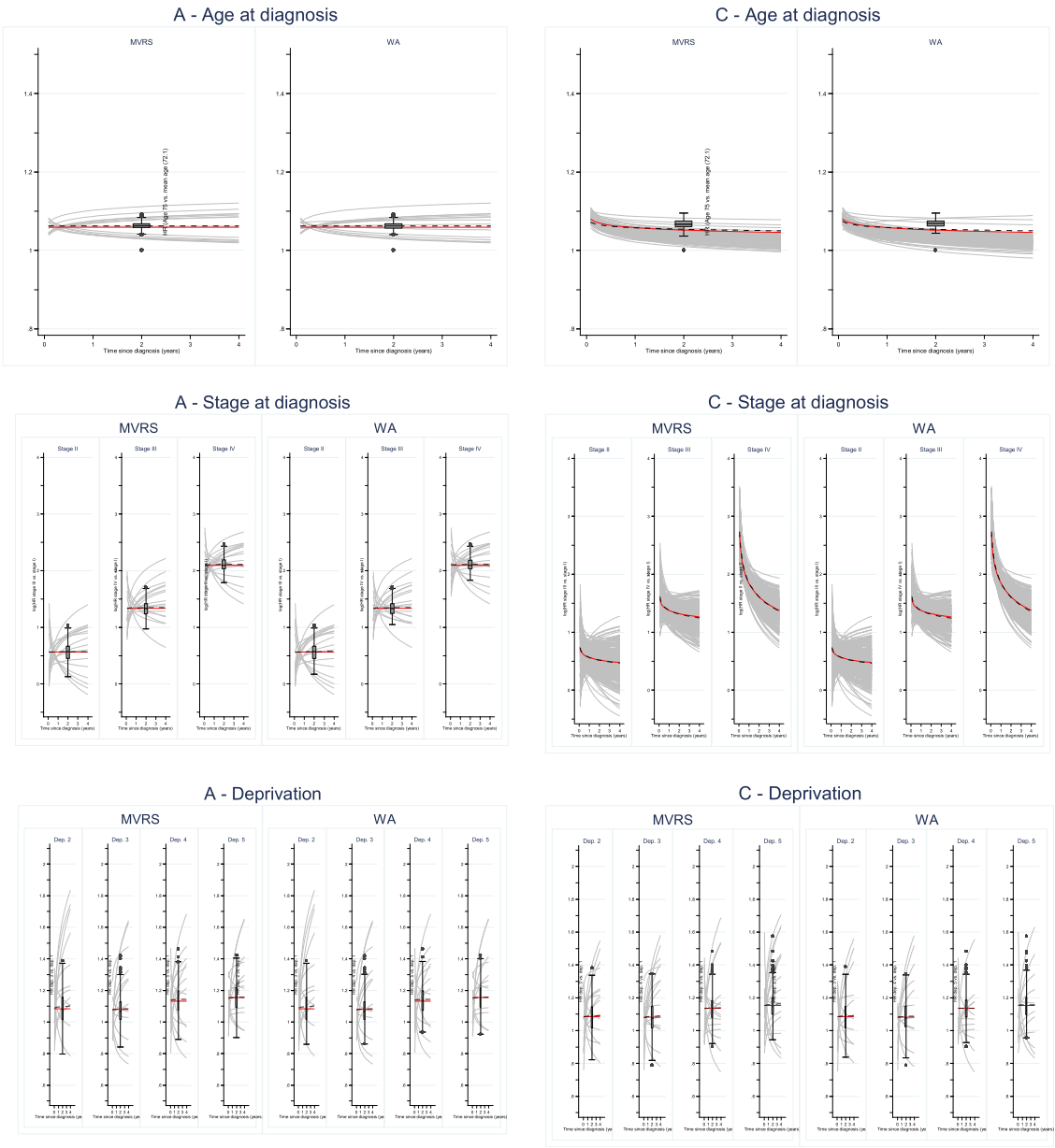


Fig. 5 Generating (red line), estimated (grey lines - time varying, box-plot - time fixed) and averaged (black dashed line) hazard ratio for age, stage, and deprivation: scenario A and C

the cohort is well captured by the models. *pABCtime* values calculated using the non-parametric estimator of net survival provides 0.3–8% higher values than for the model-based survival curves (Additional file 7).

The bias reflects the varying amount of mis-specification for each of the three algorithms. For example, higher proportions of time-dependent effect of the binary variables using W&A and aW&A lead to higher standardised bias for that variable and that algorithm (Additional file 8). The minimum in the time-varying bias is reached at around 6 months after diagnosis for all effects, when most time-dependent effects

Table 3 *pABCtime* between the mean of the individual effects or cohort net survival estimated using the selected models and the true generating effects/cohort net survival, by scenario (A-D), and model selection strategy

Cohort net survival	*Stage*	A	B	C	D	**HR age**	*Stage*	A	B	C	D
MVRS / aMVRS	I	1.67%	1.62%	0.79%	1.09%	MVRS / aMVRS	I	0.34%	2.43%	0.35%	4.40%
	II	0.94%	1.60%	0.99%	1.10%		II		3.56%		5.86%
	III	0.56%	0.53%	0.16%	1.37%		III		0.01%		0.73%
	IV	0.36%	0.36%	0.11%	0.31%		IV		0.13%		2.15%
W&A / aW&A	I	0.05%	1.63%	0.04%	0.89%	W&A / aW&A	I	0.33%	2.35%	0.23%	2.88%
	II	0.20%	1.60%	0.01%	1.00%		II		3.45%		4.13%
	III	0.06%	0.54%	0.94%	1.18%		III		0.02%		0.61%
	IV	0.13%	0.37%	0.68%	0.08%		IV		0.12%		1.26%
MFPIgen	I		0.09%		0.08%	MFPIgen	I		2.55%		2.73%
	II		0.13%		1.11%		II		3.76%		3.92%
	III		1.18%		0.99%		III		0.01%		0.63%
	IV		0.21%		0.71%		IV		0.06%		1.86%
HR stage	*Stage*	A	B	C	D	**HR deprivation**	*Deprivation*	A	B	C	D
MVRS / aMVRS	I					MVRS / aMVRS	2	1.20%	0.03%	0.34%	1.38%
	II	2.25%	3.30%	2.44%	1.01%		3	0.23%	0.03%	0.39%	0.75%
	III	1.71%	1.75%	1.20%	2.96%		4	0.93%	0.40%	0.21%	1.24%
	IV	2.49%	1.93%	2.15%	5.49%		5	0.13%	0.26%	0.53%	1.36%
W&A / aW&A	I					W&A / aW&A	2	1.20%	0.02%	0.26%	1.20%
	II	2.35%	3.36%	2.32%	1.71%		3	0.27%	0.01%	0.35%	0.57%
	III	1.66%	1.86%	0.85%	3.42%		4	0.95%	0.23%	0.12%	0.84%
	IV	2.52%	2.08%	1.80%	5.24%		5	0.13%	0.50%	0.28%	1.20%
MFPIgen	I					MFPIgen	2		0.03%		1.81%
	II		3.25%		1.39%		3		0.02%		1.02%
	III		1.67%		3.36%		4		0.44%		1.11%
	IV		1.83%		5.37%		5		0.23%		1.38%
HR comorbidity		A	B	C	D						
MVRS / aMVRS		0.31%	0.22%	0.08%	0.18%						
W&A / aW&A		0.59%	0.34%	0.12%	0.12%						
MFPIgen			0.36%		0.17%						

cross the true effect. At that point, the value reached reflects the amount of bias due to the estimated fixed effects.

Discussion

Motivated by the growing access to data on explanatory factors of cancer survival, we compared the practical use of several model selection strategies. We adapted well-recognised algorithms to the context of excess hazard models, including extensions to deal with 2-way interactions. Simulations, based on observed realistic scenarios, showed the ability of all strategies to yield proper estimation of the cohort net survival curve despite varying forms of the retained and estimated effects.

Several aspects of model selection deserve further discussion. Additionally, we aim to provide some guidelines for variable selection in the context of cancer survival epidemiology.

Subject matter knowledge

A breadth of modelling strategies exists, but very few strategies have been compared as highlighted by STRA TOS Topic Group 2 [37]. We aimed here to look at the impact that model selection strategies may have on inference based on the final selected model. Subject matter knowledge is needed all through model building, such as in decisions relative to the selection of the variables that will be tested, and the allowed forms of these variables [38], as well as how strict we are on keeping/dropping a variable or functional form. In observational studies, we

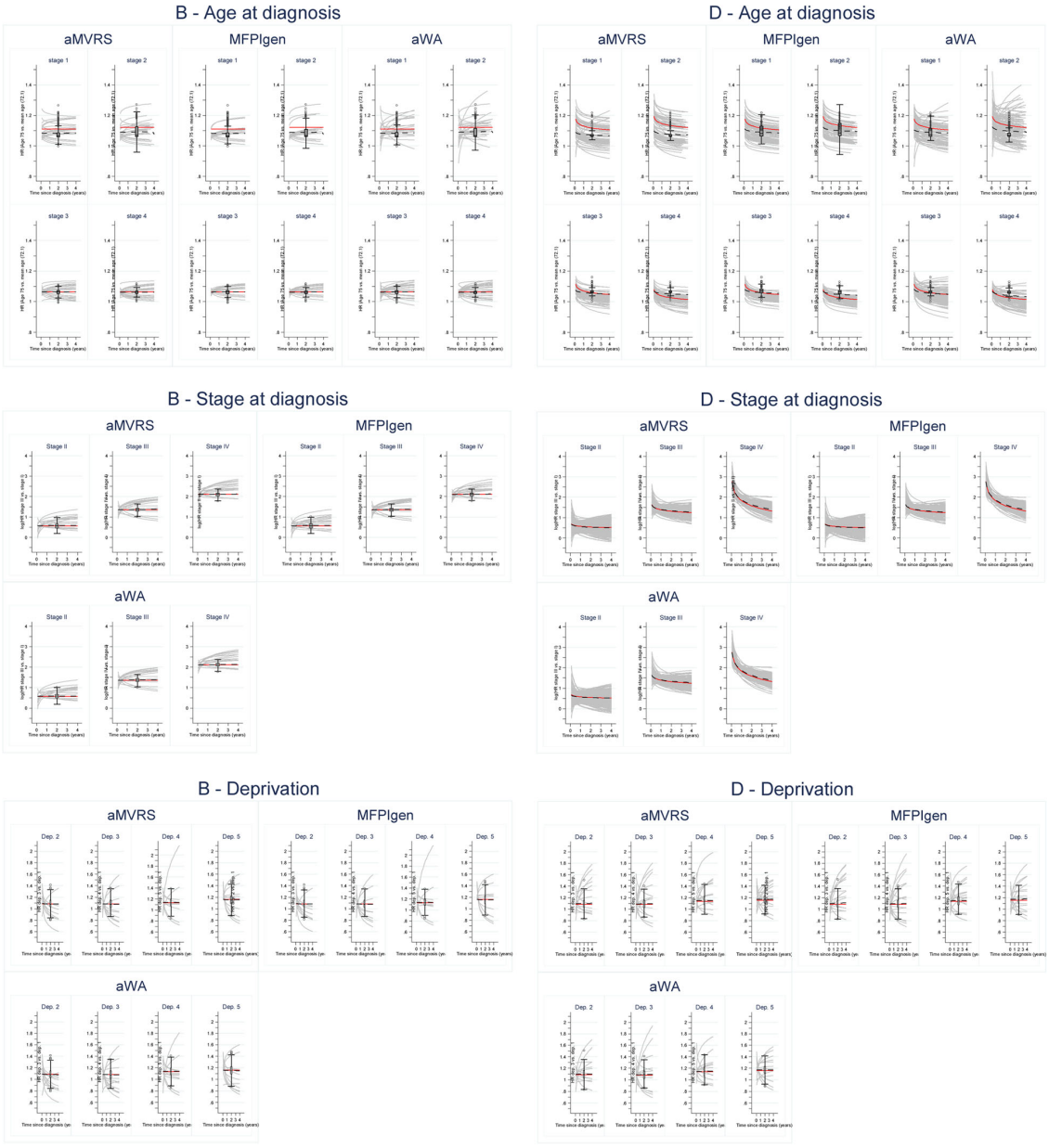


Fig. 6 Generating (red line), estimated (grey lines if estimated as time varying, or in the box-plot if estimated as time fixed) and averaged (black dashed line) hazard ratio for age, stage, and deprivation: scenario B and D

acknowledge it is almost impossible to state all aspects of a model ahead of data exploration, and model selection remains necessary. In our comparison, we concentrate on the model-building algorithms per se and assume both benefited from a similar amount of subject matter knowledge.

Time-dependent effects

A time-dependent effect is modelled if the effect of a variable, measured at diagnosis, varies with time since diagnosis, i.e. that effect is not constant with follow-up time. In the context of cancer survival, most factors such as stage at diagnosis, deprivation, emergency presentation

[39] tend to have strong effects in the months that follow the diagnosis, and these effects are likely to reduce or disappear as time passes [39]. When testing time-dependency of different factors, a long enough follow-up, as well as enough information are required to detect time-dependency.

Non-linear effects

Additionally, in order to properly assess non-linearity of the effect of a specific variable, such as age, there needs to be enough information on that variable about its own effect on the time to event: e.g. patients' age need to cover a reasonable range of all possible ages, rather than be grouped in a small part of the age distribution.

Censoring and lethality of cancer (number of events)

Lung cancer data contain relatively high proportions of events (80% 4 years after diagnosis) compared to other cancers that do not experience such high lethality. Model building strategies and variable selections are highly sensitive both to the number of events and levels of censoring. This is due to the rapidly increasing complexity of the models tested, especially when the backward-based W&A and aW&A are run. For example, in the context of lung cancer, there was non-convergence of the Stata algorithms in around 10% of the samples. Changing the starting values or running initial univariate selections did not help in reaching convergence.

It has recently been shown that 40–50 events per variable are necessary to ensure accurate estimation of coefficients [40] in the competing risk setting. In the most complex models (fitted on lung cancer) which include all interactions and time-dependent effects, i.e. 48 parameters, there was an average of 36 events per parameters in a sample of 2000 patients. When these model-building strategies were run on cancers with lower lethality, such as laryngeal cancer, with 60% censoring at 5 years, the algorithms did not converge for a larger proportion of samples, up to 20% (results not shown). In addition, after convergence, some estimated hazard ratios were unbelievably large: there was an average of only 16 events per parameters ($N = 2000$ patients) for the most complex models fitted on laryngeal cancer data.

In the relative survival data setting, a competing risk framework, competing deaths (i.e. from other causes, provided by general population life tables) are subtracted from observed events (death from any cause). This reduces further the power for detecting and retaining effects. This is not so problematic when studying lung cancer as 95% of deaths are due to lung cancer [39], i.e. 1675 lung cancer deaths among the 1765 deaths in the 2000 data samples, leading to 34 events per parameter. Less lethal cancers will see the actual numbers of cancer-related deaths be a smaller proportion of all deaths, leading to smaller number of events per variable.

Prior to running any model building strategies, we recommend that the censoring rate and the number of events are carefully examined in relation to the complexity of the models fitted. Further clinical considerations and background knowledge are helpful prior to variable selection to ensure significance tests are used with sparsity.

Sample size, model complexity

The W&A strategies tend to favour time-dependent effects and interactions, leading to complex models. This is due to the backward selection of effects. Model misspecification of some variables leads to self-confounding and confounding, which would provide wrong inference on the effects of some variables. On the other hand, the MVRS strategy leads to simpler models with additional variables wrongly selected in about 5% of models overall. However, in three out of four scenarios (B, C and D), all model selection strategies select models containing the true models in a relatively poor proportion (always below 15%). This is largely due to the size of the effects that the algorithms were trying to capture and the number of patients included in the analyses, 2000. Indeed, some effects such as non-linearity or non-proportionality of age could not be retrieved in the final selected models, due to lack of power. Releasing one or several of these small effects translates in larger proportions of models that nearly contain the generating models. More importantly, increasing the sample size to 5000 patients leads to improved detection power and higher selection proportions of the true generating model.

The adapted MVRS and W&A algorithms testing for interactions show similar properties as the original algorithms for the selection of linear/non-linear and time-dependent main effects. They show equivalent results to the MFPIgen strategy for the selection of interaction terms.

Investigating the effect of many variables of known prognostic value in a large population-based cohort of lung cancer patients, all model-building strategies lead to similar selection of effects. As expected W&A and aW&A only differed from R&S and aR&S in the shape of the effect of age, which has virtually no impact on cohort-wide net survival estimates.

Although the model-building strategies may not tend to select the same final models, and the proportion of models that do select the true generating model vary with the sample size, the number of events and the size of the effects, there is no impact on the estimation of cohort net survival, by stage at diagnosis. Estimation of cohort net survival can best be done non-parametrically as there is no assumption on the form of the association between the exposure variables and survival time. We

show that on average the model-based estimates are equivalent to the non-parametric estimates of net survival. When non-parametric estimates of cohort survival can be produced, it is good practice to use them to validate model-based estimates.

The variables whose effects are tested in the models, are only mildly correlated with a coefficient of correlation below 0.2. Another challenge in modelling non-linear effects of a variable is the potential collinearity of some spline basis (such as cubic splines). A possible solution for this, adopted here, consists of orthogonalizing the splines basis. However, high correlation between two variables may have a negative effect on the model selection strategies studied here as they are based on stepwise methods and are thus dependent on the order of testing.

Epidemiological aim of models

The ultimate aim of building exploratory models in our context is to describe variables effects on the survival experience of a cohort of cancer patients. In the simulations, the large variety of models selected by the different model-building strategies leads to varying estimations of main effects and varying levels of individual excess hazard and net survival estimates, which has implications in terms of epidemiological interpretation. Nonetheless all generated effects are well captured by the variable selection strategies, whatever their complexity. This is verified graphically and looking at the area between each estimated effect and the generated effect.

Forward-based model building strategies tend to favour simpler models, which may be a useful feature in contexts with less information (e.g. low EVP, or high censoring, or relatively small sample sizes) in order to avoid inclusion of spurious effects. Conversely, backward-based strategies tend favour more complex models, which may be useful to detect small effects in cases with larger samples and low censoring. Nonetheless, the comparison of the final models selected with different strategies may be useful in order to assess any differences on the corresponding net survival curves, and to identify potential reasons for these differences (if any) based on our previous discussion.

The strategies presented here are based on likelihood ratio tests performed in a hierarchical order. Thus, they rely on significance testing and, consequently, are prone to multiple testing as well as Type I and Type II errors. Nonetheless, all strategies let the user decide what significance level should be used for the selection of effect. We use here the conventional 5%, and test for the impact of keeping the main effects in. One could consider choosing more conservative thresholds [41] and evaluating the impact of varying thresholds on the models selected.

Model building strategy is in line with the 'data modelling culture' and is based on the idea that *a true* model generating the data does exist [42]. Although not all important variables may be available, or the true model is likely to not be among the considered models, the aim is to get as close as possible to this true model by including the relevant variables and by flexibly modelling the effect of the available ones. Shrinkage techniques (LASSO [43], Ridge, Elastic Nets [44]) could be considered, but these methods are not yet available in our relative survival context. Still in the machine learning field, methodological developments are of great interest. For example, model averaging [45] and more generally ensemble learning techniques [46] are possible avenues though interpretability of the results can be challenging, hence more appropriate outside of the descriptive modelling field.

Model selection approaches based on Information Criteria [45] (e.g. AIC and BIC) or cross-validation of the selected models, instead of likelihood ratio testing, could prove useful for selecting the proper functional forms of effects. In the context of prediction, one tends to select and use a simple model in order not to over fit the training data [47]. Following work on the topic of predictions would involve additional statistical measures for assessing predictive accuracy of the selected model for a given strategy. Measures such as discrimination and calibration would then be useful [48, 49]. However, in this work, which was mainly exploratory rather than predictive, all strategies lead to similar model-based estimates of net survival.

Large datasets and information on many factors are motivations for using complex excess hazard models. Model selection methods are essential to make sure all models are considered in a systematic fashion. Nonetheless, several aspects of the data (such as sample size, censoring, NL and TD effects) and the models (such as complexity, assumptions) deserve full consideration ahead of model selection.

Additional file

Additional file 1. Adaptation of the W&A and MVRS for testing for interactions.

Additional file 2. Comparison of model-based estimates of survival and the non-parametric Pohar Perme estimator of survival.

Additional file 3. Descriptive statistics for the sample of patients used in data simulations.

Additional file 4. Self-confounding and confounding effect due to mis-selection of variable effects (scenario A and C).

Additional file 5. Effects of age, generating (red), estimated (grey), averaged (black) for all models selected by each algorithm, scenario A-D.

Additional file 6. Effect of the extra binary variable in the models selected by each algorithm, scenario A-D.

Additional file 7. Original (red line), estimated (grey lines - time varying, box-plot - time fixed) and averaged (black line) cohort net survival. Scenario A-D.

Additional file 8. Integrated absolute estimated bias for the effects of age, stage, deprivation and the extra binary variable, by scenario (A-D) and for each model selection algorithm.

Abbreviations
AIC: Akaike Information Criteria; BIC: Bayesian Information Criteria; LASSO: Least Absolute Shrinkage and Selection Operator; MFPI: Multivariable fractional polynomial (interaction); MFPIgen: Multivariable fractional polynomial (general interaction); MFPT: Multivariable fractional polynomial (time); MVRS: Multivariable regression splines; NL: Non linear; pABCtime: Proportion of the Area Between Curves through time; PH: Proportional Hazards; TD: Time dependent; W&A: Wynant and Abrahamowicz

Acknowledgments
Not applicable.

Disclaimer
The findings and conclusions in this report are those of the authors and do not necessarily represent the views of Cancer Research UK.

Authors' contributions
CM and BR developed the concept and CM, BR, AB and FJR developed the design of the simulation study. CM was involved in the data preparation and the data linkage, carried out the data analysis and wrote the manuscript. All authors interpreted the data, drafted and critically revised the manuscript. All authors read and approved the final version of the manuscript.

Funding
This research was supported by Cancer Research UK (grant number C7923/A18525).

Availability of data and materials
The data that support the findings of this study – both in simulations and application – are available from Public Health England but restrictions apply to the availability of these data, which were used under the necessary statutory and ethical approvals for the current study, and so are not publicly available.

Ethics approval and consent to participate
Consent would not be feasible due to the large number of patients involved, of which many would be deceased. We are secondary users of this confidential population-based data and as such we hold statutory approval from the Confidentiality Advisory Group (PIAG 1–05(c)/2007) which grants us permission to process the data for the purpose of this study.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Cancer Survival Group, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. [2]Department of Mathematics, King's college, London, UK.

Received: 5 November 2018 Accepted: 6 September 2019
Published online: 20 November 2019

References
1. Sauerbrei W, Abrahamowicz M, Altman DG, le Cessie S, Carpenter J. STRengthening analytical thinking for observational studies: the STRATOS initiative. Stat Med. 2014;33(30):5413–32.
2. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC Med. 2010; 8(1):20.
3. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3(Mar):1157–82.
4. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-dimensional variable selection for survival data. J Am Stat Assoc. 2010;105(489):205–17.
5. Shmueli G. To explain or to predict. Stat Sci. 2010;25(3):289–310.
6. Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. Stat Med. 2007;26(2): 392–408.
7. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med. 2007;26(30):5512–28.
8. Wynant W, Abrahamowicz M. Flexible estimation of survival curves conditional on non-linear and time-dependent predictor effects. Stat Med. 2016;35(4):553–65.
9. Sauerbrei W, Royston P, Look M. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. Biom J. 2007;49(3):453–73.
10. Royston P, Sauerbrei W. Multivariable modeling with cubic regression splines: a principled approach. Stata J. 2007;7:45–70.
11. Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Stat Med. 2004;23(16):2509–25.
12. Sauerbrei W, Royston P, Zapien K. Detecting an interaction between treatment and a continuous covariate: a comparison of two approaches. Comput Stat Data Anal. 2007;51(8):4054–63.
13. Wynant W, Abrahamowicz M. Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. Stat Med. 2014;33(19):3318–37.
14. Esteve J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. Stat Med. 1990; 9(5):529–38.
15. Mariotto AB, Noone AM, Howlader N, Cho H, Keel GE, Garshell J, et al. Cancer survival: an overview of measures, uses, and interpretation. J Natl Cancer Inst Monogr. 2014;2014(49):145–86.
16. Belot A, Ndiaye A, Luque-Fernandez MA, Kipourou DK, Maringe C, Rubio FJ, et al. Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data. Clin Epidemiol. 2019;11:53–65.
17. Pohar Perme M, Stare J, Esteve J. On estimation in relative survival. Biometrics. 2012;68(1):113–20.
18. Pohar Perme M, Esteve J, Rachet B. Analysing population-based cancer survival - settling the controversies. BMC Cancer. 2016;16(1):933.
19. Pohar Perme M, Henderson R, Stare J. An approach to estimation in relative survival regression. Biostatistics. 2009;10(1):136–46.
20. Danieli C, Remontet L, Bossard N, Roche L, Belot A. Estimating net survival: the importance of allowing for informative censoring. Stat Med. 2012;31(8): 775–86.
21. Remontet L, Bossard N, Belot A, Estève J. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. Stat Med. 2007;26(10):2214–28.
22. Giorgi R, Abrahamowicz M, Quantin C, Bolard P, Esteve J, Gouvernet J, et al. A relative survival regression model using B-spline functions to model non-proportional hazards. Stat Med. 2003;22(17):2767–84.
23. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. Stata J. 2009;9:265–90.
24. Rubio FJ, Remontet L, Jewell NP, Belot A. On a general structure for hazard-based regression models: an application to population-based cancer research. Stat Methods Med Res. 2019;28(8):2404–17.
25. Bower H, Crowther MJ, Lambert PC. Strcs: a command for fitting flexible parametric survival models on the log-hazard scale. Stata J. 2016;16(4):989–1012.
26. Royston P, Sauerbrei W. Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables [chapter 7: interactions]. UK: Wiley; 2008.
27. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. Stat Med. 2013;32(23):4118–34.
28. Crowther MJL. P.C. Simulating complex survival data. Stata J. 2012;12(4):674–87.
29. Department for Communities and Local Government. The English indices of deprivation 2007. London; 2008.
30. Sobin LH, Gospodarowicz M, Wittekind C. TNM classification of malignant Tumours. 7th ed. New York: John Wiley & Sons; 2009.
31. Wang Z, Ma S, Zappitelli M, Parikh C, Wang C-Y, Devarajan P. Penalized count data regression with application to hospital stay after pediatric cardiac surgery. Stat Methods Med Res. 2016;25(6):2685–703.
32. Buchholz A, Sauerbrei W, Royston P. A measure for assessing functions of time-varying effects in survival analysis. Open J Stat. 2014;4:977–98.
33. Benitez-Majano S, Fowler H, Maringe C, Di Girolamo C, Rachet B. Deriving stage at diagnosis from multiple population-based sources: colorectal and lung cancer in England. Br J Cancer. 2016;115:391.

34. Elliss-Brookes L, McPhail S, Ives A, Greenslade M, Shelton J, Hiom S, et al. Routes to diagnosis for cancer – determining the patient journey using multiple routine data sets. Br J Cancer. 2012;107:1220.
35. Maringe C, Fowler H, Rachet B, Luque-Fernandez MA. Reproducibility, reliability and validity of population-based administrative health data for the assessment of cancer non-related comorbidities. PLoS One. 2017;12(3): e0172814.
36. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927;22(158):209–12.
37. Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, et al. State-of-the-art in selection of variables and functional forms in multivariable analysis -- outstanding issues 2019. Available from: https://arxiv.org/abs/1907.00786.
38. Heinze G, Wallisch C, Dunkler D. Variable selection – A review and recommendations for the practicing statistician. Biom J. 2018;60(3):431–49.
39. Maringe C, Pohar Perme M, Stare J, Rachet B. Explained variation of excess hazard models. Stat Med. 2018;37(14):2284–300.
40. Austin PC, Allignol A, Fine JP. The number of primary events per variable affects estimation of the subdistribution hazard competing risks model. J Clin Epidemiol. 2017;83:75–84.
41. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018;2(1):6–10.
42. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci. 2001;16(3):199–231.
43. Zou H. The adaptive Lasso and its Oracle properties. J Am Stat Assoc. 2006; 101(476):1418–29.
44. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B. 2005;67(2):301–20.
45. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer Science & Business Media; 2003.
46. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Data mining, inference, and prediction. 2nd ed. New York: Springer-Verlag; 2009.
47. Clayton MK, Geisser S, Jennings DE. In: Goel PK, Zellner A, editors. A comparison of several model selection procedures. New York: Elservier; 1986.
48. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Stat Med. 1999;18(17–18):2529–45.
49. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy and measuring and reducing errors. Stat Med. 1996;15(4):361–87.

Publisher's Note

1.5 Discussion

To my knowledge, this is the first independent comparison of the performances of two model-building algorithms and their adaptations to deal with the selection of interactions, in the context of excess hazard regression models. We aimed to compare the selection processes in their abilities to select predictors, their functional form, potential time-dependent effects and interactions, when modelling their effects on the excess hazard of death.

In the article, we provide guidance for the modelling of excess hazard of death, and for the best use of the data and model selection tools. This is a first step towards the selection of models relevant for the prediction and projection of cancer survival.

The application presented in the paper shows minimal impact of model-selection algorithm on the final effects selected. Indeed the final model selected by each algorithm were almost identical, with a selection of time-dependent effects for all variables, excluding deprivation. There were slight discrepancies in the selection of interactions.

This exploration of the model selection algorithms presents some limitations that would deserve further investigation. These include (i) we tested only one spline function of each continuous variable with 3dfs and did not test for simpler non-linear effects; (ii) time-varying effects are only tested as interactions between each variable and the logarithm of time since diagnosis: these effects could be more complex; (iii) we only considered interactions between a categorical and a continuous variable.

The next steps would be to explore the robustness of the model selection. For instance, we could compare the effects selected to those that would be selected in bootstrap samples of the data. Then we could also check the need for interactions by comparing the estimated parametric net survival curves to the non-parametric Pohar-Perme estimators. Our aim is to have interpretable models, to tease out the associations of interest. Generalisability is also of interest, which means we are less concerned with missing weaker effects, possible artefact to the data. Chapter 3 presents how we have extended this framework to the prediction of cancer survival.

1.6 My other contributions to the topic

I have had other exposure to excess hazard model building – and adopted a different approach –, especially while analysing and contrasting stage-specific cancer survival in six high-income countries, as part of the International Cancer Benchmarking Partnership. [95] At the time, we performed model selection from a set of pre-specified models chosen

based on background knowledge and data availability. Such an approach was also adopted by the SUDCAN partnership, [96] when estimating survival for all cancers in England and Wales, [97] and when looking at variations in access to surgery for patients with colorectal cancer in four high-income countries. [98]

Maringe C, Walters S, Butler J, Coleman MP, Hacker N, Hanna L, et al. Stage at diagnosis and ovarian cancer survival: evidence from the International Cancer Benchmarking Partnership. Gynecologic oncology. 2012 Oct;127(1):75-82

Walters S, *Maringe C*, Butler J, Rachet B, Barrett-Lee P, Bergh J, et al. Breast cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK, 2000-2007: a population-based study. Br J Cancer. 2013 Mar 19;108(5):1195-208

Maringe C, Walters S, Rachet B, Butler J, Fields T, Finan P, et al. Stage at diagnosis and colorectal cancer survival in six high-income countries: a population-based study of patients diagnosed during 2000-2007. Acta Oncol. 2013 Jun;52(5):919-32

Walters S, *Maringe C*, Coleman MP, Peake MD, Butler J, Young N, et al. Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004-2007. Thorax. 2013 Jun;68(6):551-64

Benitez Majano S, Di Girolamo C, Rachet B, *Maringe C*, Guren MG, Glimelius B, et al. Surgical treatment and survival from colorectal cancer in Denmark, England, Norway, and Sweden: a population-based study. The Lancet Oncology. 2019;20(1):74-87

Chapter 2

Tools for evaluating predictions

2.1 Introduction

We have seen in the previous Chapters that non-parametric and parametric estimates of net survival can be derived from population-based cancer registrations (see Setting). Chapter 1 reflects on practical and algorithmic aspects relative to the choice of models and functional forms for the effects of explanatory factors on the excess hazard of death. These models may be used to make prediction of survival.

Indeed, long-term cancer survival for patients most recently diagnosed with cancer and for whom follow-up is not yet recorded or available cannot be estimated directly from their data. We will assume that information on historic cohorts of cancer patients are available and can bring useful information for estimating survival for more recent cohorts. For example, the effect of their socio-demographic and tumour characteristics on their excess hazard of death informs predictions of excess hazard for newly-diagnosed patients. We rely on flexible regression models to synthesise past associations between prognostic factors and outcome, and to predict future outcomes.

In this Chapter we detail the statistical tools available to characterise the overall performance, validity and accuracy of predictions. We focus on the tools developed for, or adapted to, the time-to-event and competing risks settings. We present a measure of explained variation for excess hazard modelling. Finally, we discuss how other standard tools are and could be modified to accommodate the specificities of the relative survival data setting.

2.1.1 Prediction models around us

There are areas where predictions are key: weather forecast, finance and economy such as stock market prediction or economic growth. Prediction models are also popular in demography, where population growth is of interest. [99–101] Actually it extends to areas such as bacteriology, ethnology, zoology, [102] or medicine where 'populations' refer to bacteria, social groups, animals, or the interest is in the future incidence of disease or the speed of tumour growth. The general public is used to predictions of life expectancy from life tables. [103, 104] In the medical setting, predictions of incidence and mortality patterns [105, 106] are useful for healthcare planning purposes. Additionally the creation of risk scores (or scoring rules) [107, 108] for developing a disease or a complication are often proposed to identify early high-risk patients or as an aid for diagnosis. The three main approaches to forecasting described in Chambers [109] are still relevant: qualitative techniques, time series analysis and projection and causal models.

Many scoring rules have been developed to aid clinical decision. [110] Despite their popularity among researchers, these rules are not used routinely by clinicians for a number of reasons, including: poor generalisability, requirement of a complex data collection, absence of guidelines on their use. Guidelines have been published to foster a transparent reporting of multivariable prediction models, to minimise risks of bias and maximise clinical interest and use. [22, 23, 111]

2.1.2 Time-to-event data

Survival models differ from other forms of regression models since the outcome is the knowledge of two variables, length of follow-up (T) and vital status at the end of the follow up (δ). Due to the presence of a time dimension, there is both loss to follow-up and censoring (C). This means that for some patients the event of interest is not observed during the follow-up time, hence the time to event will be unknown. Survival models make use of all records, including censored ones, to estimate the association of prognostic factors (X) with the outcome.

Although loosely referred to as 'survival models', most models are fitted on the hazard scale. [32] In other words, the association between prognostic factors and the rate of occurrence of an event in the sample is modelled through follow-up time: $\lambda(T|X) = f(t, X)$. Hazard-based regression models process raw information on patients' survival time and vital status to estimate a fluctuating hazard function with time. Other models exist such as accelerated failure time models, on the survival scale, [112] or the general hazard model. [113]

In the context of survival data for patients diagnosed with a given disease, predictions from survival models can be sought to create a risk score for new patients. Patients are classified given their baseline socio-demographic, or disease-related characteristics and a survival probability is derived based on their risk score. Risk-score development provides clinical aid to decision making and helps in communicating their prognosis to patients. These types of predictions are in-sample predictions (see Glossary), as patients characteristics are within the range of values used for training the model. Besides such predictions tend to be made for patients with specific characteristics, and as such represent individual predictions.

We aim to perform out-of-sample prediction of cohort survival, in other words allowing prediction of cancer survival for patients (and groups of patients) who do not contribute to model building. We will explore if and how the assessment tools developed in the context of in-sample prediction may be useful for devising the best predictive model or set of models, given our data. [114]

There are tools for the assessment of prognostic models in the context of time-to-event data. [114–124] Such tools have often been adapted from the general linear regression framework to account for censoring that comes with following patients through time. Time is a key feature that influences the prognostic ability of a model, such as fluctuating performance at different times after the start of follow up.

2.1.3 Survival models: what can be predicted?

Due to data complexity, there are several quantities that can be predicted from survival models. These include (a) individual probabilities that the event occurs at a given time, (b) individual probabilities that the event occurs before a given time, or (c) prediction of individual survival times. One can also measure (d) the instantaneous or (e) cumulative force of mortality at given times or its fluctuations through time. Finally, (f) the ratio or (g) differences in hazard or survival values can also be estimated at different times after diagnosis. Quantities (a) and (b) are defined on the probability scale, while (c) is on the time scale, and (d) and (e) are on the hazard scale.

Predictions of remaining survival time are often a source of interest in patients with terminal disease and their carers. Nonetheless it was shown by Henderson and colleagues that these point estimates of time carry poor predictive capability; [122] this is widely quoted and recognised. [116, 119] They acknowledge the interest in survival time prediction, especially in communication to patients, but insist they are provided with a measure of confidence in the values, and further cautionary words. A measure of the model's overall performance can provide such caution.

There are not many tools for checking the accuracy of predictions developed in the specific context of excess hazard models in the relative survival data setting. Contrary to other fields, in which outcomes are fairly well defined (binary outcomes: occurrence of a disease, relapse, death...; continuous outcomes: costs, score ...), there are multiple quantities of interest in survival analyses, depending on the research question or the audience. [2] Therefore, one needs to choose the quantities of interest as well as a most appropriate measure of predictive accuracy. Different statistical measures of predictive accuracy show complementary facets of the quality of predictions. We aim to highlight in this Chapter if and how the tools developed in the overall survival context could be used in the relative survival data setting.

2.1.4 Prediction and projection from multi-variable models

Prediction and projection flow naturally from multivariable modelling. Providing one is clear on the assumptions behind the estimations of association and trends (see Chapter 1), it is computationally easy to extrapolate the identified trends and use the estimated parameters to (i) predict longer-term outcomes for patients whose early follow-up contributed to model-building or (ii) project patient outcomes for those who did not contribute to model building at all.

In this Chapter, we leave aside the complexities of model selection: we assume a series of models have been evaluated, and a specific model is considered fit for predicting the outcome of interest. We are interested in the evaluation of the predictions from the selected model. Identifying a *final* model is a topic that is covered in depth in Chapters 1 and 3.

2.2 Evaluation of predictive models

Predictive performance measures are typically presented in the following categories, [110, 125] and are commonly used for the evaluation of risk prediction models: [126–128]

Overall performance: it estimates the distance between observed and predicted outcomes, using loss functions such as the Brier score. Measures of explained variation are other types of overall performance measures. They quantify the amount of variation in outcomes explained by the prognostic variables that constitute the model. The most famous of these measures is the coefficient of determination, R^2, defined in linear regression. Many statistics have been proposed in the survival analysis field to provide a tool equivalent to the R^2. [129, 130]

Calibration: it assesses the ability of the model to predict accurately the absolute risk for groups of patients. Measures of calibration such as the calibration plot, and general goodness of fit statistics, evaluate the agreement between observed outcomes and predictions.

Discrimination: the discrimination measures highlight the ability of the model to separate observations into risk groups that is to rank individuals from low to high risk. These include sensitivity and specificity, the Area Under the receiver operating Curve (AUC), and the concordance statistic (c-index).

These statistics were developed outside of the time-to-event context, and some have been adapted to deal with censoring and sometimes competing risks setting. There is general agreement on the difficulties of making predictions in the survival field, [131, 132] mostly relative to: inadequate models, sampling variability, lack of explanatory power of the survival model, problems extrapolating to new data. [132]

In the field of survival analyses, despite wide use of prognostic classification models, they have 'rarely [been] subjected to a rigorous examination of their adequacy' highlighted Graf et al in 1999, [116] and then Schoop et al in 2008 stated "there exist[ed] as yet no standard approach to assess the predictive accuracy of [survival] models". [119] This is in stark contrast with the well-established measures used in binary or continuous outcomes setting. There are however recent reviews to address the issue of lack of reporting on the quality of prediction models and model-based predictions. [22, 111] Additionally a specific topic group on 'Evaluating diagnostic tests and prediction models' (topic group 6) is part of the STraTOS initiative. [133, 134] In the relative survival data setting, the use of excess hazard models is expanding, but very few manuscripts mention model selection, model testing and model performance. Nonetheless, 'rigorous examination' is still key, as the statements highlighted above also apply to the context of disease-specific survival models. I briefly review how measures of model evaluation have overcome the challenges of the competing risks framework. Then I open the discussion to ideas on how to adapt to the challenges of the relative survival data setting.

2.3 Measures of overall performance

Overall performance measures aim at summarising the distance between the predicted and observed outcomes, typically using loss functions. Measures of explained variation also provide an indication of the overall explanatory power of a model. In linear regression, loss functions correspond to the error or residuals. For the evaluation of survival models they should be adapted to the predicted outcome of interest:

'If one is interested in estimating the hazard, the loss function should ideally involve the hazard; if interest is in the survival function, the loss function should involve the survival function.' [52]

The down sides of relying on a loss function to characterise the predictive accuracy of models include: (1) censored observations complicate the calculations of the loss functions, although some corrections may be used [116, 118] and (2) they are outcome-dependent. Given the different outputs available from survival models, many loss functions have been proposed. We will concentrate here on describing in details the Brier score, as it has been adapted to accommodate for time-to-event data, as well as the competing risks setting through the prediction error.

Let us introduce two patients, A and B with a description of what happens to them post-diagnosis (Figure 2.1). Patient A experiences the event of interest at time t_A and B survives beyond the end of follow up t_{final} and is therefore censored alive at t_{final}. For $X = A$ or B, let us suppose we also have $\hat{T}_X$, their predicted survival time and $\hat{S}_X(t^*)$ their predicted probability to survive beyond t^*, such that $t^* \leq t_{final}$. We define $Y_X(t) = \mathbb{1}_{T_X > t}$, the at-risk indicator for patient X, and $dN_X(t) = \mathbb{1}_{T_X = t}$ an indicator that the event for patient X happens at time t. Figure 2.1 contrasts the predicted survival curves throughout follow up for patients A and B (plain lines), together with what was actually observed for these patients in the form of their at-risk indicator Y_X (dashed lines).

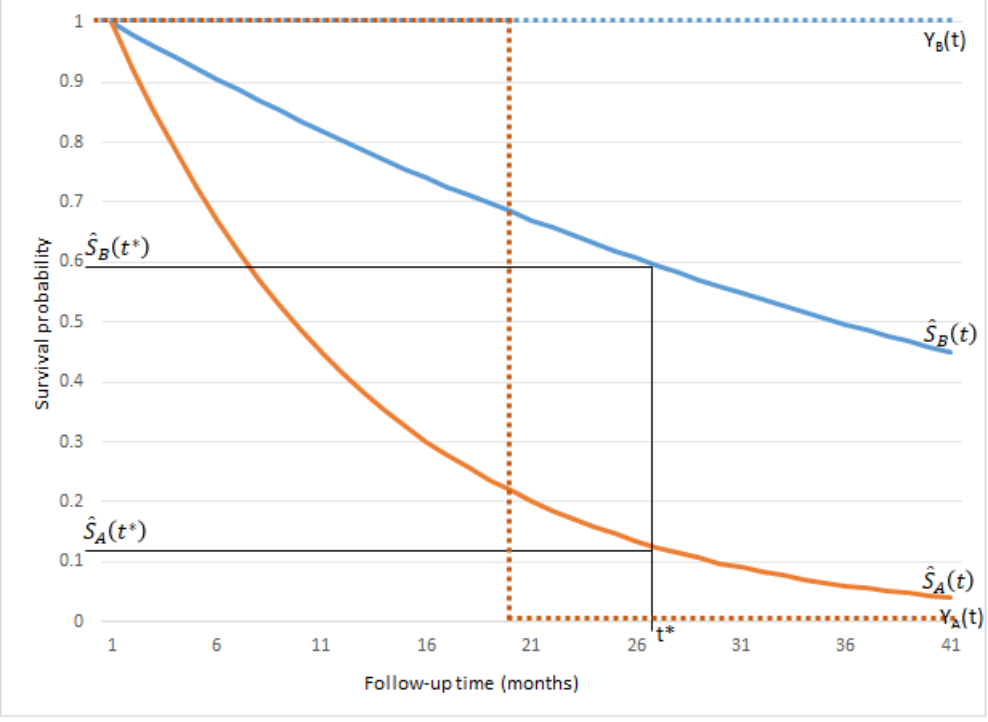


Figure 2.1: Observed and estimated survival functions for fictive patients A and B

2.3.1 The Brier score

The Brier score is a loss function, based on a quadratic loss contrasting observed and predicted outcomes. In survival, the Brier score contrasts the values of Y_X and $\hat{S}_X$.

Original score

The Brier score [135] is a measure of the expected loss in using the predicted outcome values in lieu of the observed outcome. It is calculated as a mean square error between observed and predicted outcomes. The decision space for the Brier score is between 0 and 1. Values closer to zero reflect smaller error and better overall performance.

The formula of the Brier score for a sample of N observations is

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(t) - \hat{S}_i(t)\right)^2. \tag{2.1}$$

Time-to-event data

Due to possible censoring of the follow-up time, not all $Y(t)$ may be observed and a weighting is proposed [116] assuming that censored patients can be adequately represented by patients with complete information. The weights assigned to fully observed patients are given by the pooled Kaplan Meier estimator, $\hat{G}(t) = \hat{p}(C > t)$ of the censoring event, C:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\mathbb{1}_{(\tilde{T}_i \le t, \delta_i = 1)}}{\hat{G}(\tilde{T}_i)}\left(Y_i(t) - \hat{S}_i(t)\right)^2 + \frac{\mathbb{1}_{\tilde{T}_i \ge t}}{\hat{G}(t)}\left(Y_i(t) - \hat{S}_i(t)\right)^2\right]. \tag{2.2}$$

where $\tilde{T} = min(T, C)$ with C the censoring time, and $\delta = 0$ or 1 the censoring indicator with $\delta = 1$ for uncensored (fully observed) patients.

In 2.2 we see the contribution that patients who experience the event prior to t ($\mathbb{1}_{\tilde{T}_i \le t, \delta_i = 1} = 1$) make to the Brier score: $\left(0 - \hat{S}_i(t)\right)^2$. This is weighted by the inverse of the probability to be uncensored by the time of event $\tilde{T}_i$. The contribution of patients still in the risk set at time t, ($\mathbb{1}_{\tilde{T}_i \ge t} = 1$) is $\left(1 - \hat{S}_i(t)\right)^2$, weighted by the inverse of the probability to be uncensored at the horizon time t. For both weights, if the probability of no censoring beyond $\tilde{T}_i$ or t is high, the weights tend to 1; if it is low, meaning many have been censored prior to $\tilde{T}_i$ or t, the weight is over 1. Censored patients prior to time t only contribute to the calculation of $BS(t)$ through the weights, as both $\mathbb{1}_{\tilde{T}_i \ge t} = 0$ and $\mathbb{1}_{\tilde{T}_i \le t, \delta_i = 1} = 0$ for them.

Since $BS(t)$ is function of t, integrated versions of the Brier score can be calculated at any time t^* when there are patients at risk: $\int_0^{t^*} BS(t)dt$.

The closer the $BS(t)$ is to zero the better the prognostic accuracy of the model on which it is based. Further research demonstrated the consistency of this mean-squared error of prediction. [118] Adaptations of the Brier score to deal with time-dependent covariates representing updated information through follow-up time are also proposed using conditioning on patients still at risk at the time at which the updated information is available. [119, 136]

Extension to competing-risks models

Since the relative survival data setting is part of the competing risks setting, we further focus here on developments of the Brier score (or prediction error) to deal with competing risks, proposed by Schoop et al. [137] They consider two competing events, and assume one is interested in the prediction error for one of these events, say event 1. To that end, they modify equation 2.1, for the Brier score, to that of a prediction error (PE), such that we now look at the individual probabilities of experiencing the event of interest by time t: $p(T_i \leq t, \delta_i^1 = 1)$, rather than the event free survival time $p(T_i > t)$.

$$PE(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{1}_{(T_i \leq t, \delta_i^1 = 1)} - \hat{\pi}_i^1(t|Z_i)\right]^2. \tag{2.3}$$

Where $\hat{\pi}_i^1(t|Z_i) = \hat{p}(T_i \leq t, \delta_i^1 = 1))$ is the predicted cumulative incidence function for event 1, and $\mathbb{1}_{(T_i \leq t, \delta_i^1 = 1)}$ is the indicator variable that event 1 happens before time t, for patient i with covariables Z_i.

Due to censoring, Schoop et al. propose to weight the score by the individual probabilities to be a complete case, that is uncensored by time t: either have experienced any type of event before t, $\mathbb{1}_{(\tilde{T}_i \leq t, \delta_i^1 = 1 \text{ or } \delta_i^2 = 1)}$, or still be at risk of an event at t, $\mathbb{1}_{(\tilde{T}_i \geq t)}$. [137] This corresponds to the same weights as those defined above and introduced in equation 2.2. A consistent estimator for the prediction error is therefore defined as

$$PE(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{1}_{(\tilde{T}_i \leq t, \delta_i^1 = 1)} - \hat{\pi}_i^1(t|Z_i)\right]^2 * w^1(t, \tilde{T}_i, \delta_i, \hat{G}(t), Z_i). \tag{2.4}$$

with $w^1(t, \tilde{T}_i, \delta_i, \hat{G}(t), Z_i) = \frac{\mathbb{1}_{(\tilde{T}_i \leq t, \delta_i \neq 0)}}{\hat{G}(\tilde{T}_i|Z_i)} + \frac{\mathbb{1}_{\tilde{T}_i \geq t}}{\hat{G}(t|Z_i)}$.

More recently, Wu and Li propose another weighted estimator for the Brier score, in the context of censored time-to-event competing risks analyses. [138] Using information from uncensored records at a given time t, a case status is imputed to censored patients weighted by the conditional probability of being a case by time t. Such weights require the estimation of cumulative incidence function and survival probabilities from all at risk at t. These weights are also proposed for estimating sensitivity and specificity for the ROC curves.

2.3.2 Explained variation

Measures of explained variation report the amount of variation in the outcome explained by the modelling of the explanatory factors. The squared coefficient of correlation, R^2, is the measure of explained variation most frequently used in general linear models. There are a range of measures of explained variation proposed in the survival setting. [129] The properties of such measures have been assessed in the specific context of time-to-event data, such as their performance under the effect of censoring, their interpretation, and their dependence on transformation of timescale. [129] The proportion of variation in survival explained by a model complements the information derived from the parameter estimates and their associated p-values in understanding the relative importance of the factors in explaining the levels of survival. It is recommended to provide a measure of explained variation, along with model-based measures of effects and p-values. Such a measure will show how much (or often how little) predictive value a model has, despite sometimes highly significant prognostic factors. [129]

A little like the numerous loss functions available, there are many views on what should be the relevant characteristics of a measure of explained variation. Hence many measures of explained variation, fitting with sets of desirable criteria exist. Schemper and Stare [129] classify the measures based on three definitions of the R^2 in linear regression. They compare the properties of the measures, in particular how the measures react to administrative censoring at end of follow-up t^*, and to transformations of time. Independence of the measure of explained variation to administrative censoring is only possible if one assumes that a model observed on time $[0, t^*]$ is valid beyond t^*, implying extrapolation of the model. [139] Such measures are grouped in three classes in Choodari-Oskooei [140, 141]: explained variation, explained randomness and predictive accuracy. In this section we concentrate on explained variation measures, indicating how much variation in the outcome is explained by a set of explanatory variables. Explained randomness refers to the precision with which we estimate the underlying process that generated the data. [141] Predictive accuracy correspond to measures of loss, akin to the Brier score, detailed in section 2.3.1. All three of these classes of statistics are informative in the context of cancer survival. The concept of explained variation seems particularly informative as a measure of how well one might expect to predict cancer survival in future cohorts of patients.

Stare et al. [142] proposed a measure of explained variation, RE, to satisfy a list of criteria that none of the existing measures were fulfilling. These criteria aim to make this measure useful for models with time-varying or time-dependent covariates and effects, for parametric and semi-parametric models, and which have an interpretation independent of the model fitted. [142] The new proposal is based on ranks. At each time of event, all predicted risk

scores are ranked, and the rank of the patient who fails is compared to the average rank of patients still at risk (the null model, i.e. a rank that corresponds to no information on who is most at risk). The information provided by the model (the distance between predicted and null model ranks) is then divided by the overall distance that needs to be explained (the distance between the null model rank and the observed rank, 1). That proposal is intuitive, and easily estimated and interpreted. Its values range between -1 and 1, with 0 being 'no variation in ranks is explained by the model'. The properties of RE make it a perfect candidate to be adapted to competing risks models in which a specific event might be of interest (see section 2.6.1).

2.4 Calibration

Calibration refers to the agreement between predicted and observed outcomes for different groups of the entire cohort. Such agreement is more likely when the sample of data are large enough for variable and model selection, and over-fitting is contained. We describe below some of the tools to assess the agreement between observed and predicted outcomes, internally and on external datasets.

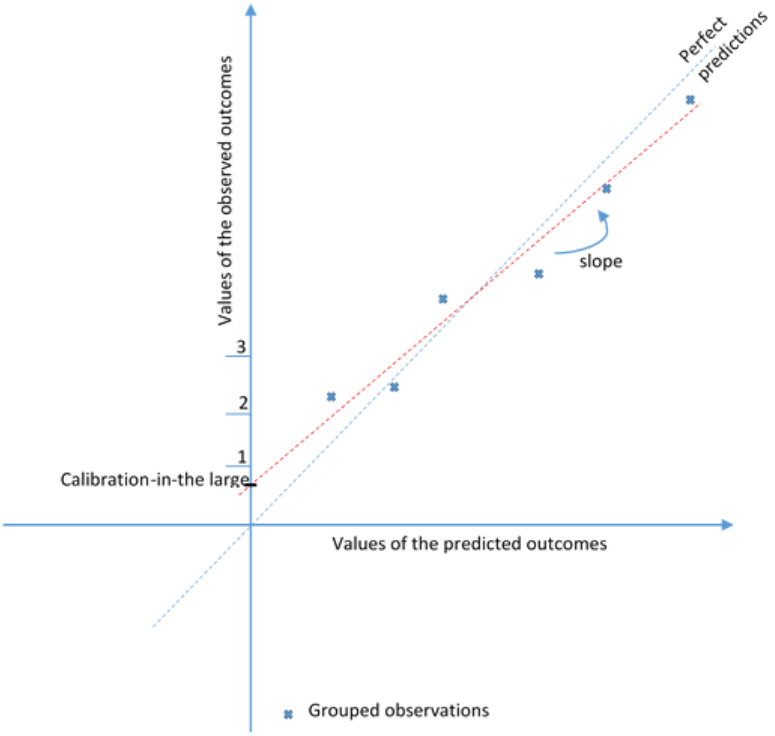


Figure 2.2: Calibration plot of actual observed outcomes *vs.* predicted outcomes for a hypothetical model

2.4.1 The Calibration plot

Checking the calibration of a model gives insights on whether a given prognosis model is also fit and consistent when run on another set of data. For example, the calibration plot graphically compares predicted and observed outcomes and provides tools for recalibrating the model to the new data available. Perfect calibration is found when all data points lie on the first diagonal. The calibration plot provide measures of apparent calibration when it is done on the data on which the model's parameters are estimated.

In the time-to-event context, the calibration plot is presented at a time horizon, and predicted survival probabilities can be compared to observed survival probabilities for different groups of patients. The groups must be defined in such a way that the estimation of survival is robust.

A regression line through the scattered values of the calibration plot can be estimated:

i its intercept will reflect the systematic bias when making predictions in a new sample (calibration-in-the-large). It represents the systematic difference between the mean of the predicted outcomes and the mean of the observed outcomes, in the new data;

In the training data, as well as in bootstrap samples of that data, there should be perfect correspondence between the average of the predicted and observed outcomes, by construction. [143]

ii the slope of the regression line between predictions and observations should approach 1 to suggest perfect calibration. [144]

In the training data we expect the slope to be 1. At internal validation, if the slope is smaller than 1, it shows the amount of shrinkage required for the model parameters to be better calibrated to new patients.

2.4.2 The Wally plot

The Wally plot is another type of calibration plot. It stems from the following definition of calibration: a model is well calibrated if, among all subjects with a predicted risk of x% at a given time t, we find x% of them who experience the event, based on the law of large numbers. [145] The Wally plot was first proposed to check the fit of linear regression models to assess calibration, while ruling out random sampling as a cause of mediocre calibration.

Similarly to the calibration plot, it makes sense to construct the Wally plot on a sample of patients who did not contribute to estimating the parameters of the risk prediction

model. Patients are split in G risk groups of approximately equal sample sizes. For any given risk group, two bars are plotted contrasting the observed proportions of patients who experienced the event of interest (the y-axis of the calibration plot), those for whom $\delta_i = 1$, to the predicted risk, $\hat{\pi}_i$, as defined from the model-based predictions (the x-axis of the calibration plot). For all group g of size N_g with $g = 1...G$, we have:

$Obs_g(t) = \frac{1}{N_g} \sum_{i=1}^{N_g} \delta_i(t)$

and

$Pred_g(t) = \frac{1}{N_g} \sum_{i=1}^{N_g} \hat{\pi}_i(t)$.

Blanche et al. [145] suggest inverse-probability-of-censoring weighting to deal with right censoring, with weights calculated as the inverse of the Kaplan-Meier estimate of the censoring survival function. The plot is related to the Hosmer-Lemeshow test for a departure from the calibration assumption using a chi-square distribution. [146] That test was recently criticised on the basis that 'it is based on artificially grouping patients into risk strata, gives a P value that is uninformative with respect to the type and extent of miscalibration, and suffers from low statistical power'. [147]

With the Wally plot, a visual inspection of the discrepancy between the bars informs the assessment of calibration for the specific risk prediction model and data. To ease the decision relative to the quality of calibration, Blanche et al. propose to create a series of such calibration plots, obtained from similar data simulated under the calibration assumption. All graphs displayed together in random order form the Wally plot. If the 'real' graph does not stand out, it reflects good calibration of the model to the data.

2.5 Discrimination

Measuring discriminative ability corresponds to determining the capacity for a multivariable model as a whole, or additional predictors in a model, to separate observations into well-defined risk groups. We borrow tools from the framework of binary responses, in which a test (positive or negative) has to inform on disease status (case or control).

2.5.1 Sensitivity and Specificity

These measures are well defined in the framework of binary response models (such as logistic and probit regressions) and are used to qualify the ability of a test (such as a clinical test or tool, a regression model, physical examination...) to correctly identify

cases and controls. Tests and disease status are both performed and contrasted at a given point in time. Sensitivity (Se) measures the capacity, for a given test M, and a chosen threshold c, to classify a case as a case, while specificity (Sp) indicates how well that test at that threshold classifies a control as such.

Nonetheless disease status and disease outcomes are time-dependent. It naturally flows that it would be of interest to assess how well a prognostic marker, M, measured at baseline – cancer diagnosis in our setting –, can differentiate patients based on their predicted outcomes at a given point in time, during the available follow-up.

If $D(t)$ is the disease status of a given patient at time t, Heagerty et al [148] proposed the following definitions of time-varying sensitivity and specificity measures, for given thresholds c:

$$se(c,t) = p(M > c|D(t) = 1)$$
$$sp(c,t) = p(M < c|D(t) = 0)$$

These proposals are extended with 'incident/cumulative sensitivity' and 'static/dynamic specificity' measures, [149] which reflect how cases and controls may be considered through time. Cases can be defined as *incident*, when subject i is a case at the only time t such that $T_i = t$, or cases are *cumulative*, when we consider subject i a case at all times t that verify $T_i \leq t$. Similarly controls can be *static*, such that they contain subjects with a survival time longer than a pre-defined time t^*, or *dynamic*, and defined at each time t, as subjects with $T_i \geq t$. The definitions, with corresponding sensitivity and specificity, are summarised in table 2.1, and illustrated in Figure 2.3 for the fictive patients A (case) and B (control).

Table 2.1: Measures of sensitivity and specificity for time-to-event data with censoring

	Cases	**Measures of sensitivity**
Incident	$T_i = t, dN_i^*(t) = 1$	$\mathbb{I}Se(c,t) = p(M_i > c\|T_i = t)$
Cumulative	$T_i \leq t, N_i^*(t) = 1$	$\mathbb{C}Se(c,t) = p(M_i > c\|T_i \leq t)$
	Controls	**Measures of specificity**
Static	$T_i > t^*, t^* \; fixed$	$\overline{\mathbb{D}}Sp(c,t) = p(M_i \leq c\|T_i > t^*)$
Dynamic	$T_i > t, \forall t$	$\mathbb{D}Sp(c,t) = p(M_i \leq c\|T_i > t)$

We define $M_i = X_i^T\beta$, the outcome of the prognostic model we aim to evaluate, $dN_i^*(t) = N_i^*(t) - N_i^*(t-) = \mathbb{1}_{T_i=t}$ is the counting process, and c the threshold for deciding on vital status. With the formulation of incident cases and dynamic controls, models that allow for effects of longitudinal variables and time-varying factors can be analysed and assessed. Using these definitions, the interest can be in correct classification of patients still at risk

at a horizon time t, using updated markers, measured later than at time 0 in the follow-up. [149] Alternatively, cumulative cases and dynamic controls are appropriate to evaluate the prediction accuracy of a marker measured at baseline to distinguish between subjects who experience an event before t and those who do not, and therefore define the high-risk population. The estimation of the sensitivity and specificity involve estimation of the joint distribution of the marker and the survival. This is estimated using non-parametric nearest neighbour estimation techniques. [148, 149] The basic principles are that the estimations are done based on a set of subjects with the closest value of the marker.

There are variations of these time-varying measures of sensitivity and specificity fit for the competing risks setting. [150] In that setting, when $T_i = t$ or $dN_i^*(t) = 1$ we need to specify further what the cause of failure j is, such that, for instance $\mathbb{I}se(c,t) = p(M_i > c | dN_i^*(t) = 1, \delta = j)$.

Components of the measures of sensitivity and specificity above are estimated using Kaplan Meier [148] or Nearest Neighbour Estimation. [138, 148, 150, 151]

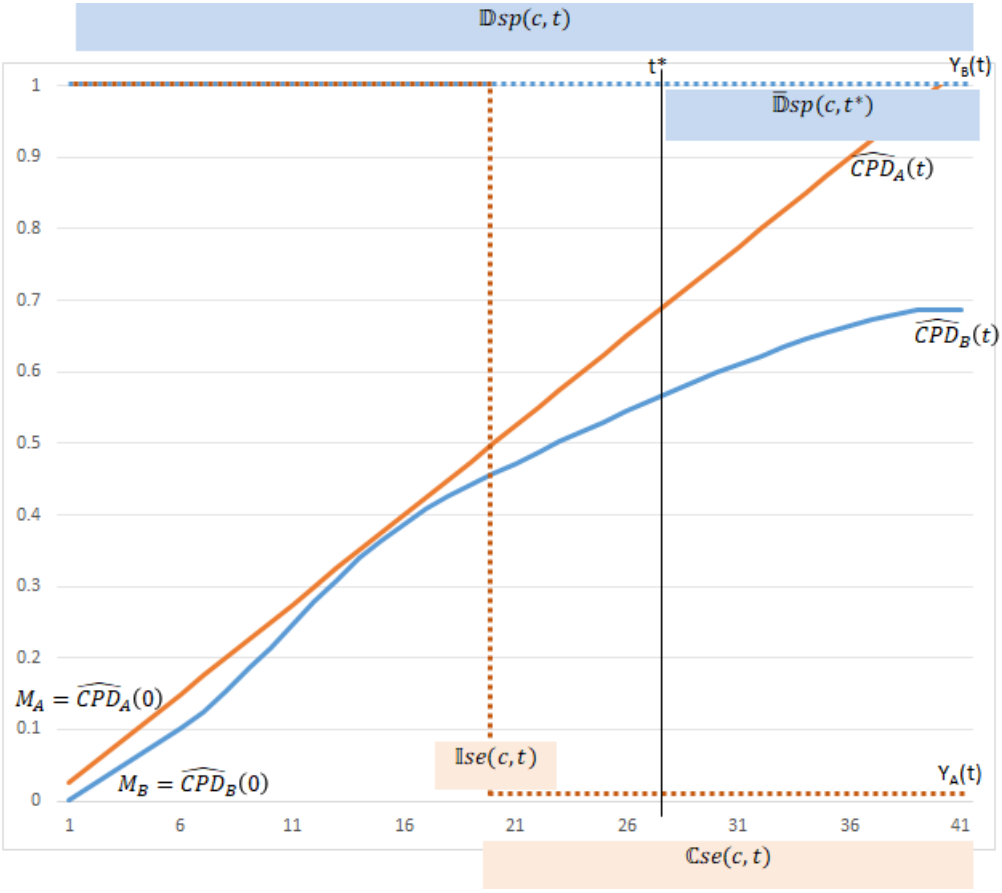


Figure 2.3: Example of sensitivity and specificity for patients A and B.

CPD stands for crude probability of death, [2] a quantity of interest from excess hazard models introduced in Chapter 1, that could be used as a marker.

2.5.2 The ROC plot

The ROC (Receiver Operating Curve) curve is the plot of the sensitivity (true positive rate) against the false positive rate, i.e. 1-specificity with different levels of the cut-offs defining the groups (Figure 2.4). On the plot is also displayed the 45-degree line, corresponding to no discriminative ability of the model. Any line above the first diagonal indicates increasing discrimination, with 1 reflecting a model that perfectly discriminates cases from controls. The ROC curve informs on the prognostic potential of a model. [149] It can also be useful to compare the discriminatory power of different models.

In the time-to-event field, we briefly saw in 2.5.1 that new definitions of sensitivity and specificity measures, and subsequent ROC curve, stem from (1) different definitions of cases and controls, (2) when and how often the risk score is measured, (3) inclusion of time varying effects, or (4) presence of censoring or competing events. [152]

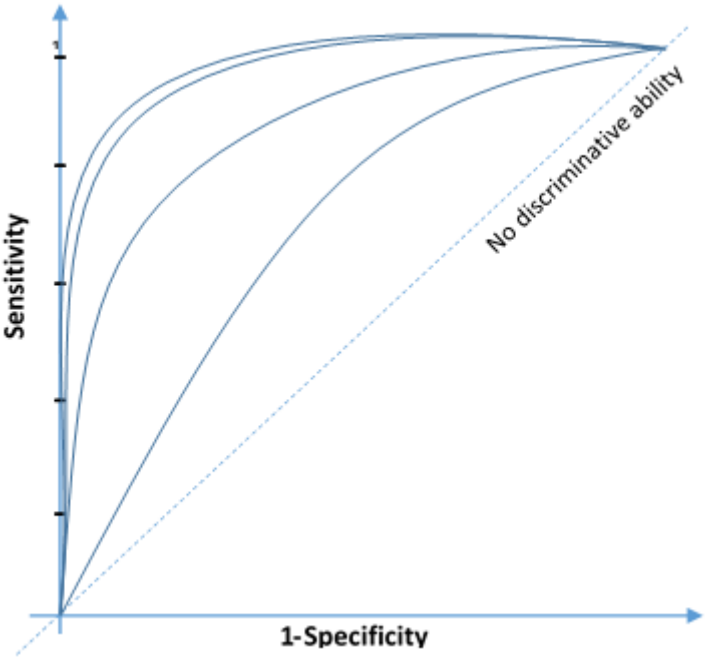


Figure 2.4: ROC plot for 4 hypothetical prediction models

2.5.3 Area Under the Curve and C-statistic

Measuring the area under the ROC curve (AUC) summarises the information contained in the shape of the curve into a single number. AUC is also referred to as the C-statistic or the C-index, its generalisation to survival data. [75] The AUC represents a measure of concordance between the disease or survival status and the marker or risk score derived from a model. [149] It represents the probability that the risk score of a randomly selected

case (or patient who experiences the event of interest) is higher than the risk score for a randomly selected control (or surviving patient).

On one hand the ROC can be the basis for the calculation of the AUC, using standard numerical integration. On the other hand one can look at the proportion of pairs for which the risk score of the subject that failed is higher than the risk score of the subject still at risk (C-index). The estimation and interpretation of the C-index was first extended from the logistic regression models to survival models by Pencina and d'Agostino. [153] For two patients i and j for whom their actual survival times T_i and T_j are as such that $T_i \neq T_j$, and either both or at least one of them experience the event of interest before the end of follow up, we can write π_C probability of concordant pairs, and π_D, probability of discordant pairs as follow:

$$\pi_C = p\left(T_i < T_j \text{ and } \hat{T}_i < \hat{T}_j \text{ or } T_i > T_j \text{ and } \hat{T}_i > \hat{T}_j\right)$$
$$\pi_D = p\left(T_i < T_j \text{ and } \hat{T}_i > \hat{T}_j \text{ or } T_i > T_j \text{ and } \hat{T}_i < \hat{T}_j\right)$$

$\hat{T}_i$ and $\hat{T}_j$ are the predicted survival times, estimated as the expected mean survival time. The proportion of concordant pairs is then defined as $C = \pi_C/(\pi_C + \pi_D)$.

The C-index is a measure of how well a risk prediction model discriminates between groups of patients defined by their outcome (deceased/alive). However such a measure cannot describe how well a given model is at predicting individual risks.

In the competing risks setting, Wolbers et al [154] propose a cause-specific concordance index, based on the cumulative incidence function of the event of interest, and related to the time-varying AUC introduced above. [150, 155] Similarly to other measures, they offer an inverse-probability-of-censoring weighting to account for right censoring. The proposed measure $C_1(t)$ correctly ranks events of interest (events of type 1) up to time t, and discriminate them from competing events (events of other type). $C_1(t)$ is defined as follows: $C_1(t) = p\left(M(t, X_i) > M(t, X_j) | \delta_i = 1 \text{ and } T_i < t \text{ and } (T_i < T_j \text{ or } \delta_j = 2)\right)$

2.6 Validation measures in the relative survival data setting

In this section, we present ideas for extending some of these measures into the relative survival data setting. The specificities of that setting mean that one cannot use these statistics directly, despite their corrections and weights offered for dealing with competing risks. We focus on extending a measure of overall performance, RE, given its broader application, even outside the field of prediction. This measure also complements the information provided by model parameters and their p-values. The original intention was

also to use the measure of explained variation as a tool for model selection for prediction. Nonetheless, due to its insensitivity to parametrisation (see details in the manuscript in 2.6.1), we realise it will not be directly useful for model selection.

As possible extensions of this work, we believe the approach adopted for the *RE* could be useful for adapting other statistical tools such as the Brier score, or measures of sensitivity and specificity to the relative survival setting. These adapted measures would characterise further the predictive properties of excess hazard models. We present avenues for further research on this topic in subsections 2.6.2-2.6.4.

2.6.1 Explained variation of excess hazard models, Maringe et al., Statistics in Medicine, 2017

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	273792	**Title**	Mrs
First Name(s)	Camille		
Surname/Family Name	Maringe		
Thesis Title	On the prediction and projection of cancer survival		
Primary Supervisor	Prof. Bernard Rachet		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Statistics in Medicine		
When was the work published?	06 April 2018		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I was lead author on the paper. I designed the study, and simulations, in consultation with co-authors, I performed the analyses, interpreted the results and presented the results in the manuscript. I drafted the manuscripts and received comments from all co-authors.

SECTION E

Student Signature	[redacted]
Date	06-02-2020

Supervisor Signature	[redacted]
Date	5 February 2020

Received: 21 June 2017 | Revised: 30 January 2018 | Accepted: 31 January 2018

DOI: 10.1002/sim.7645

RESEARCH ARTICLE

WILEY Statistics in Medicine

Explained variation of excess hazard models

Camille Maringe[1] | Maja Pohar Perme[2] | Janez Stare[2] | Bernard Rachet[1]

[1] Cancer Survival Group, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

[2] Department of Biostatistics and Medical Informatics, University of Llubljana, Vrazov trg 2, SI-1000 Ljubljana, Slovenia

Correspondence
Camille Maringe, Cancer Survival Group, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.
Email: camille.maringe@lshtm.ac.uk

Funding information
CRUK, Grant/Award Number: C7923/A18525

The availability of longstanding collection of detailed cancer patient information makes multivariable modelling of cancer-specific hazard of death appealing. We propose to report variation in survival explained by each variable that constitutes these models. We adapted the ranks explained (RE) measure to the relative survival data setting, ie, when competing risks of death are accounted for through life tables from the general population. RE is calculated at each event time. We introduce weights for each death reflecting its probability to be a cancer death. RE varies between −1 and +1 and can be reported at given times in the follow-up and as a time-varying measure from diagnosis onward. We present an application for patients diagnosed with colon or lung cancer in England. The RE measure shows reasonable properties and is comparable in both relative and cause-specific settings. One year after diagnosis, RE for the most complex excess hazard models reaches 0.56, 95% CI: 0.54 to 0.58 (0.58 95% CI: 0.56–0.60) and 0.69, 95% CI: 0.68 to 0.70 (0.67, 95% CI: 0.66–0.69) for lung and colon cancer men (women), respectively. Stage at diagnosis accounts for 12.4% (10.8%) of the overall variation in survival among lung cancer patients whereas it carries 61.8% (53.5%) of the survival variation in colon cancer patients. Variables other than performance status for lung cancer (10%) contribute very little to the overall explained variation. The proportion of the variation in survival explained by key prognostic factors is a crucial information toward understanding the mechanisms underpinning cancer survival. The time-varying RE provides insights into patterns of influence for strong predictors.

KEYWORDS
excess hazard models, explained variation

1 | INTRODUCTION

Complex, multivariable modelling of time-to-event data is easily accessible through user-friendly specific commands in common statistical software.[1-4] In such models, the effects of prognostic factors on hazard of death are modelled and estimated. Possible non-linearity and time dependence of their effects can be incorporated. The model gives the usual estimates of effect and *P*-values, but often the estimation of survival for the cohort is the metric of choice.

Datasets in population-based research contain information on virtually all patients in a given area or country for a given period of time: it can represent such large numbers that statistical significance does not bring much information

on the relative importance of prognostic factors. A measure of explained variation does not aim at providing information on how well a model fits the data at hand but provides information on how much of the variation in survival between records is explained by the model, and hence by the prognostic factors that compose the model.

Although survival models do not carry good prediction properties, there is a number of measures proposed for evaluating their prognostic characteristics[5] by ways of measures of prediction accuracy,[6] discrimination potential,[7] and the proportion of variation explained.[8] Most measures have been designed in the context of the Cox model,[9] widely used in traditional survival analyses or clinical trials. However, when focussing on survival from a disease, eg, cancer, survival analysis needs to account for competing risks of death. In the population-based cancer survival context, the exact cause of death of patients is unknown or considered unreliable. In this context, we rely on the relative survival data setting, in which the hazard of death from the cancer, or excess hazard, is estimated by comparing the overall mortality of the cancer patients to their expected mortality provided by life tables built for the general population from which the cancer patients come.[10] The effects of explanatory variables on the excess hazard can be modelled using various excess hazard models.[3,11] Net survival, the survival of the cohort of cancer patients, cancer being the only cause of death, can be derived from such models, providing these are well specified. The assumption of informative censoring is replaced by a more plausible assumption of independence of the forces of mortality, providing the effects of the variables stratifying the life tables, such as sex, age, region, deprivation, and ethnicity, are adjusted for in the model.[12]

In this paper, we adapt a measure of explained variation, ranks explained (RE),[13] to the context of excess hazard models in the relative survival data setting. We address challenges related to the specificities of that setting and the excess hazard modelling, while the interpretation of the adapted RE is kept as simple as with the original RE measure. This is exemplified by an extensive illustration using population-based cancer registry data on patients diagnosed with colon or lung cancer in England.

The next section summarises the characteristics of the measure of explained variation, RE, then presents the excess hazard models and how RE was adapted to the relative survival data setting. In a third section, we describe the design of our simulation-based analyses aimed at exploring the features of RE. The following section presents an application based on colon and lung cancer patients in England. The discussion wraps up the main advantages and limitations of the measure proposed.

2 | METHODS

2.1 | The RE measure in the overall survival setting

The RE measure, standing for "ranks explained", was introduced by Stare et al.[13] It aims at providing a measure of the variation in the ranks observed in survival-time data explained by a given model. It can be viewed as a generalisation of the C-index.[14] It satisfies the following list of criteria:

(1) Applicability to multiple end-point survival
(2) Facility to incorporate time-varying and/or dynamic covariates and/or time-dependent effects
(3) Model-free interpretation on a well-understood scale, to allow comparison between non-nested models
(4) Applicability to both parametric and semiparametric models
(5) Consistency under general independent censoring mechanisms, including intermittent missingness and delayed entry or truncation

Some of these points, particularly (2), (3), and (5), make the measure appealing to the excess hazard context.

Technically, the sum of the variation in ranks, explained by the model is compared with the sum of the total variation in ranks there is to explain. The "unit" is the rank that each record is given at each failure time t_i, ie, the predicted position at which the record under observation will fail among all observations that have yet to fail (observations in the *risk set* R_i). The total variation is viewed as the difference between the ranks allocated under a "null model" ($r_{i,null}$), and the ranks allocated under a "perfect model" ($r_{i,perfect}$), ie, the record that fails is always given rank 1:

$$r_{i,null} = \frac{(k+1)}{2} \qquad \forall k \in R_i, \text{ i.e. } t_k > t_i$$
$$r_{i,perfect} = 1.$$

We define the "null model" as a model in which all records that have not yet failed are given the same mean rank: it corresponds to a scenario in which one would lack information regarding the expected time to failure of the individuals in the risk set, and all individuals would therefore have the same probability to fail next.

The variation that is explained by a proposed model is the difference between the ranks allocated under a "null model" and the ranks that are allocated under the proposed model ($r_{i,model}$).

$$r_{i,model} = 1 + \sum_{k \in R_i} \mathbb{I}_{\lambda_{k(t_i)} > \lambda_{i(t_i)}}$$

Where $\lambda_{k(t_i)}$ and $\lambda_{i(t_i)}$ are the hazards for patients k and i, respectively, at patient's i time of failure t_i.

The final statistic sums these differences over all individual failure times so that the statistic is defined, in the case of single-event survival data by:

$$RE = \frac{\sum_i \left(r_{i,null} - r_{i,model}\right)}{\sum_i \left(r_{i,null} - r_{i,perfect}\right)}. \tag{1}$$

Through censoring patients leave the cohort. In order for those who stay in the cohort to be representative of those who left, we weight records that are more likely to have missing observed failure time. Typically, the weights are the reverse Kaplan Meier estimates ($\frac{1}{\widehat{G_{it}}}$), in the case of survival data with right censoring.[5,15] The delta method is also used to provide a formulation for the variance of RE. Full details can be found in Stare et al.[13]

$$RE = \frac{\sum_i \int_0^\tau \frac{1}{\widehat{G_{it}}} * \left(r_{i,null}(t) - r_{i,model}(t)\right) dN_i(t)}{\sum_i \int_0^\tau \frac{1}{\widehat{G_{it}}} * \left(r_{i,null}(t) - r_{i,perfect}(t)\right) dN_i(t)} \tag{2}$$

In Equations 1 and 2, the sum is by default over all observations N that fail in the sample. It is also of interest to estimate instantaneous measures of explained variation, termed *local RE*, for which the sum is made over the x records that fail around each successive observed failure times throughout the entire follow-up. The value of x depends on the cancer, but the illustrations presented here used a window of 20 failures.

$$localRE = \frac{\sum_i \int_{t-x/2}^{t+x/2} \frac{1}{\widehat{G_{it}}} * \left(r_{i,null}(t) - r_{i,model}(t)\right) dN_i(t)}{\sum_i \int_{t-x/2}^{t+x/2} \frac{1}{\widehat{G_{it}}} * \left(r_{i,null}(t) - r_{i,perfect}(t)\right) dN_i(t)} \tag{3}$$

2.2 | The excess hazard model

Net survival is the survival that would be observed in our population of cancer patients, had cancer been the only possible cause of death.[16] Net survival can be estimated in the cause-specific setting or in the relative survival setting. The main difference between the 2 settings is the knowledge of the cause of death.

In the cause-specific setting, the exact cause of death is known, and the failure indicator reflects whether the patient dies from his/her cancer (failure is coded 1), did not die (failure is 0), or died from a cause other than cancer (failure is 0 or 2). It is straightforward to adapt RE to cause-specific survival models: the only difference is that RE is evaluated at each cancer death rather than each death (see Figure 1A).

In the relative survival setting, cause of death is not available or not deemed reliable; therefore, population life tables are used in the modelling of excess mortality to adjust for mortality due to other causes, also termed expected or background mortality. Population life-tables reflect the pattern of survival of the general population, from which the cancer patients are drawn. In population-based cancer survival, the relative survival setting is the setting of choice for the estimation of net survival through excess hazard modelling.

We aim that RE gives a measure of how much of the *cancer* survival variation observed between individuals is explained by a specific excess hazard model: we remove the impact of other causes and isolate the effects of potential additional variables on cancer mortality.

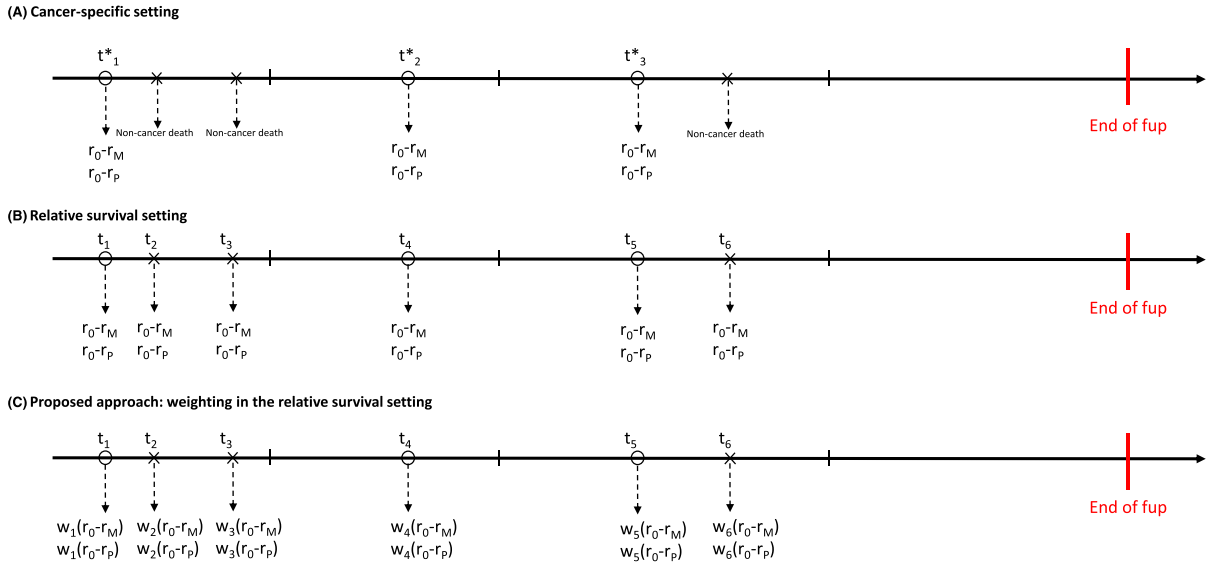


FIGURE 1 Calculation of RE in different settings (A) Cancer-specific setting (B) Relative survival setting (C) Proposed approach: Weighting in the relative survival setting. ◯ time of cancer death; X time of non-cancer death; | time of censoring; r_M: Rank as estimated from the model-derived hazard of death; r_0: Average rank of the records in the risk set; r_P: 1; w_i: probability of cancer death [Colour figure can be viewed at wileyonlinelibrary.com]

2.3 | RE measured from an excess hazard model

(a) Weights

In the same way that consistent estimators of net survival can be obtained in both relative survival and cause-specific settings, we want RE calculated in both settings to agree. In the cause-specific setting, RE is evaluated only at times of cancer deaths (Figure 1A). By contrast, the relative survival setting uses all failure times regardless the cause of death (Figure 1B) for which RE needs to be adjusted (Figure 1C).

Therefore, we propose to weight each event time with quantities reflecting the probability that the event is happening due to the cause of interest at the time considered. We therefore consider

$$w_i = p(dN_{E_i}(t) = 1 |\, dN_i(t) = 1) \tag{4}$$

where $N_{E_i}(t)$ is the counting process associated with the cause of interest, and $N_i(t)$ is the all-cause counting process.

We define the weights as the ratio of the excess mortality due to cancer λ_E, over the sum of the excess and expected (population, λ_P) mortality.[17] Both hazards are estimated at the time of death.

$$w_i = w(t_i) = \frac{\lambda_{E_i}(t_i)}{\lambda_{E_i}(t_i) + \lambda_{P_i}(t_i)} \tag{5}$$

Take the practical example of the cause-specific setting: if we were to use weights, differences in ranks would be evaluated at times at which patients are censored due to death from other causes, but their weight, hence contribution, would be 0, because the probability that the event is a cancer death is null. To mirror this in the relative survival setting, weights would tend to 0 when the probability of cancer death is highly unlikely, and weights would tend to 1 when the probability of cancer death is highly likely.

We want to show that the total number of cancer events can be estimated by the sum of weights w_i. By law of total probabilities, we have

$$p(dN_{E_i}(t) = 1) = p(dN_{E_i}(t) = 1 |\, dN_i(t) = 1) * p(dN_i(t) = 1) + p(dN_{E_i}(t) = 1 |\, dN_i(t) = 0) * p(dN_i(t) = 0) \tag{6}$$

Because $p(dN_{E_i}(t) = 1 | dN_i(t) = 0) = 0$, and dN variables are binomial variables, if one sums Equation 6 over individuals and event times, after changing the order of summation and expectation, one gets:

$$\mathbb{E}\left(\sum_{i=1}^{n}\sum_{t_k} dN_{E_i}(t_k)\right) = \mathbb{E}\left(\sum_{i=1}^{n}\sum_{t_k} w_i * dN_i(t_k)\right) \tag{7}$$

Given that $dN_i(t_k) = 1$ if and only if $t_i = t_k$, Equation 7 can be written as follows:

$$\mathbb{E}\left(\sum_{i=1}^{n}\sum_{t_k} dN_{E_i}(t_k)\right) = \sum_{i=1}^{n} w_i \tag{8}$$

The total number of cancer events can thus be estimated by the sum of weights: depending on the quality of the approximation of the expected mortality hazard by the general population life tables and the excess hazard model to estimate cancer-specific mortality, the sum of weights will approach the number of cancer deaths.

We define RE for excess hazard models, REw, as follows:

$$REw = \frac{\sum_i \int_0^\tau \frac{w_{it}}{\widehat{G_{it}}} * \left(r_{i,null}(t) - r_{i,model}(t)\right) dN_i(t)}{\sum_i \int_0^\tau \frac{w_{it}}{\widehat{G_{it}}} * \left(r_{i,null}(t) - r_{i,perfect}(t)\right) dN_i(t)} \tag{9}$$

(b) Null models

In order to adapt RE to the relative survival setting, we kept the null model defined in Stare et al[13] and presented in Section 2.1 above; additionally the use of weights reflects the probability that an event is the event of interest.

Nonetheless, alternative null models have been considered, which assume some features of the excess hazard model a "given". For instance, we tested a null model that conveyed the life table information. The "null" rank ($r_{i,null}$) attributed to each patient at each event time t_i was derived from decreasing expected (population) mortality rates measured at t_i.

$$r_{i,null} = \text{rank}(\lambda_{P_i}(t_i))$$

It meant that for RE to be large, the effects of variables such as age, present in both the population life tables and the excess (cancer) hazard model, would need to have a different effect on the expected hazard and on the excess hazard. For example, age has a strong effect on both the expected mortality and the excess mortality; hence, both $r_{i,null}$ and $r_{i,model}$, respectively, are close to 1 for most patients i. Therefore, the individual difference $r_{i,null} - r_{i,model}$ will be slightly positive only when $r_{i,null} > r_{i,model}$, ie, when the effect of age on the expected hazard is smaller than the effect of age on excess mortality. A large β_{age} in the excess hazard model can therefore lead to a small overall RE: a result that is hard to interpret. Similarly, because some factors can cease to be discriminant for cancer survival years after diagnosis, the individual differences $r_{i,null} - r_{i,model}$ become very negative so the local RE and even the overall RE could reach very negative values.

We also tested a null model that integrated the additive structure of the overall mortality into excess and expected hazards.

$$r_{i,null} = \text{rank}(\lambda_{P_i}(t_i) + \lambda_0(t_i))$$

Nonetheless, defining a model which only contained that structure with no further assumption was challenging, and was confusing the interpretation of RE.

We believe the null model presented and used in Stare et al[13] in conjunction with our weighing remains the most relevant approach for the adaptation of the original RE to excess hazard models. Hence, $r_{i,null}$ represents the mean rank of all observations in the risk set at time t_i, reflecting a complete absence of knowledge on what observation will fail next. In this way, RE estimated through cause-specific or relative survival settings using weights will have the same interpretation.

Several outputs can be defined from the explained variation measure:

(a) Time-varying REw, REw(t):

a. it is considered as a function of follow-up time and reports the values of REw cumulated up to given times
b. REw is the cumulative measure calculated over the entire follow-up.

This is the main measure together with its variance or confidence interval.

(b) Local REw, an instantaneous measure of REw, measured using events happening between 2 pre-defined times, possibly moving through the follow-up.

This measure is exploratory, designed to investigate further specific explained variation patterns. It is advised to report smoothed curves of the local instantaneous REw values and time-varying REw(t).

3 | SIMULATIONS

We performed simulation studies to understand the properties of REw defined in the context of excess hazard models. The simulations also demonstrate the characteristics of REw such as the information it brings over the usual model outputs and how sensitive REw is to model mis-specification.

3.1 | Simulation strategy

(a) Data

We used information on 5809 breast (women) and 2418 lung (men and women) cancer patients diagnosed in England in 2000 with a valid stage at diagnosis. The potential maximum follow-up was 8 years for each patient, to the 31st December 2007, and information on their age, deprivation status, and stage at diagnosis was available. Due to passive follow-up, no censoring happens prior to the end of follow-up. Breast and lung cancers were chosen for their differing death patterns: 93% of lung cancer patients vs 30% of breast cancer patients die in the 8 years following diagnosis, and cancer deaths account for nearly 95% and around 60% of all deaths in lung and breast cancer respectively.

(b) Expected survival times

Expected survival times were simulated by extracting expected mortality rates, λ_P, from sex-specific, age-specific, year-specific, and deprivation-specific life tables, defined at each month of age and every calendar month. Moving forward, at each anniversary day of diagnosis, patient records were merged to these life tables in order to get a patient-specific expected mortality rate λ_P for that exact day. The survival time u, simulated for each patient from an exponential distribution with mean λ_P, was compared with 1 month to determine the expected survival time: if u was always greater than 1, the patient over-lived every month and was still alive at the end of the 8-year follow up. The failure indicator equals to 1 when the subject dies (whatever the cause) or 0 otherwise.

(c) Parameters of the simulations

Fully parametric models were fitted on the log cumulative hazard scale[1,11] to model the excess hazard of death using the STATA command, stpm2. Model-based information, such as the parameters of the baseline log-cumulative excess hazard, and the estimated effect parameters, was used to simulate a thousand survival times (outcome) for each of the 5809 breast and 2418 lung cancer patients. We kept the original values of the patients' sex, age, deprivation, and stage at diagnosis (observed covariate distribution). The aim of these simulations is that the simulated survival times resemble realistically observed survival patterns (see annex).

(d) Cancer survival times

We designed 2 simulation scenarios: a *simple* one, S1 only containing linear proportional effects of age at diagnosis, and a more *complex* scenario, S2, with non-proportional and non-linear effects of age, and non-proportional effects of categorical stage and deprivation (see Box A).

Survival times for S1 were simulated according to the following function for the log cumulative excess hazard:

$$\ln(H_{S1}(t; age)) = \ln(H_0(t)) + \beta_{age} * age$$

with $\ln(H_0(t)) = s(\ln(t); \gamma)$, s being a non-orthogonalised restricted cubic splines function of ln(t), with up to 3 degrees of freedom, placed at tertiles of the distribution of times.

Survival times for S2 were simulated according to the following function for the log cumulative excess hazard:

$$\ln(H_{S2}(t; age; stage; deprivation)) = \ln(H_0(t)) + f_{age}(age) * (1 + \ln(t)) + \sum_{i=2,3,4} (\beta_{stage_i} * stage_i + \alpha_{stage_i} * \ln(t) * stage_i) + \sum_{i=2,3,4,5} (\beta_{dep_i} * dep_i + \alpha_{dep_i} * \ln(t) * dep_i)$$

with $\ln(H_0(t)) = s(\ln(t); \gamma)$, s being a non-orthogonalised, restricted cubic splines function of ln(t) with up to 3 degrees of freedom, placed at tertiles of the distribution of times.

A general algorithm involving numerical integration and root-finding techniques generated the cancer-specific survival times from these complex parametric distributions.[18] We used the survsim command implemented in STATA.[19]

Overall survival time is the minimum between cancer-specific survival times, as simulated in S1 or S2, expected survival times derived from population life tables and the maximum follow-up time (8 years). From each simulated dataset, we retained the simulated expected, cancer and overall survival times, and the corresponding vital status indicators.

To make sure our simulated excess hazard and survival curves are realistic, we compared them to the original real-life hazard and survival curves (Figure 1 in Annex). More details are provided in the tables of bias and coverage (Annex) for both scenarios S1 and S2.

Because the process that generated the survival times is known, it is straightforward to assess the properties of REw calculated in several different estimation models. The estimation models M1 and M2 are well-specified as they include the same variable structure and form of effects than the simulation scenarios S1 and S2, respectively. The other models M3 to M10 are mis-specified because simulation and estimation models differ (see Box A).

We expect 1000 simulated datasets to be sufficient to offer a good overview of the properties of REw. All models were fitted on each of the 1000 simulated datasets for S1 and S2, and REw, REw(t), and local REw were calculated and their values retained for the assessment of their properties.

Excess hazard models and cause-specific hazard models both estimate the same quantity: an estimate of net survival can be derived from both strategies when cause of death is reliably known. Similar agreement is therefore expected between the values of RE measured in cause-specific and REw in relative survival settings.

3.2 | REw—Weighting system

Each individual contribution to REw was weighted by the probability that the event represents a death from cancer. The sum of these weights over all patients who died is an estimate of the number of cancer deaths in the population. Figure 2 compares the actual number of cancer deaths to the sum of weights, for each of the 2 simulation scenarios S1 and S2 for breast and lung cancers.

Over the 8 years of follow-up, there were on average 1070 breast cancer deaths among the breast cancer patients, ie, 18.4% of patients with breast cancer representing around 60% of deaths; and on average 2159 lung cancer deaths in patients with lung cancer, ie, 90% of patients representing 95% of deaths. Over the 1000 datasets simulated in each of the 2 scenarios, the sum of the weights, used in the calculation of REw, agreed with the actual number of cancer deaths, used in the cause-specific setting (Figure 2).

The agreement between REw values obtained from relative and RE in cause-specific approaches was nearly perfect, both in simulation scenarios S1 and S2 for breast and lung cancers, and at 1, 5 and 8 years after diagnosis (Figure 3). The larger variability observed in the breast cancer plots was expected and shows the greater instability of the excess hazard models due to the smaller portion that the breast cancer deaths represents among all deaths in that population (60%), contrasting with the burden of lung cancer deaths in lung cancer patients (95%).

We explored a critical scenario in which cancer mortality is very low compared with all-cause mortality: we selected stage I to II breast cancer patients aged 70 to 99 years at diagnosis. In that sample, REw was still behaving properly despite weights that were slightly over-estimated. That over-estimation can have an increasing or decreasing impact on REw depending on the directions of the effects of factors included in both the life table and the excess hazard model.

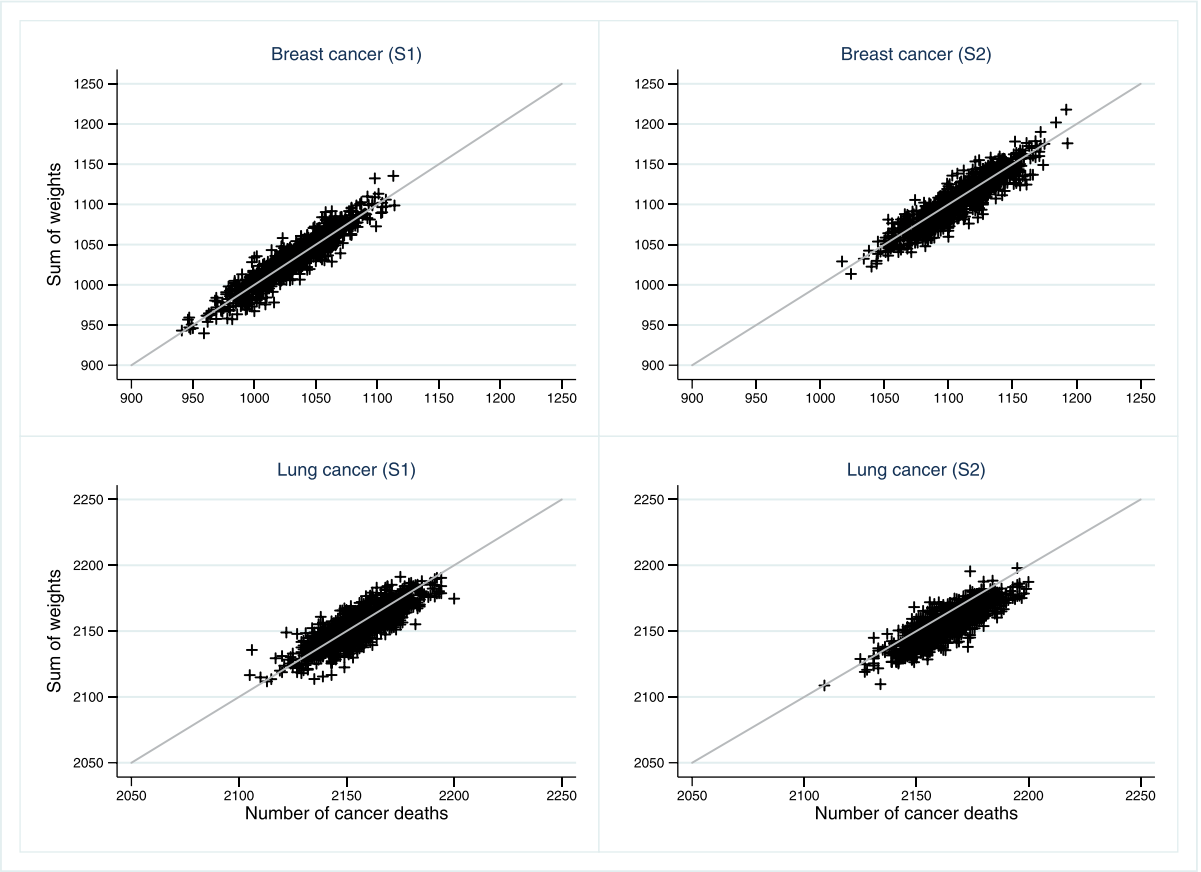


FIGURE 2 Sum of weights and actual number of cancer deaths, for each of a 1000 simulated datasets, by cancer and simulation scenario. S1: Simulation scenario 1, linear proportional effect of age at diagnosis; S2: Simulation scenario 2, non-linear non-proportional effect of age, non-proportional effects of categorical stage and deprivation [Colour figure can be viewed at wileyonlinelibrary.com]

Box A. Simulation and estimation scenarios

S1	**Simulation scenario 1, linear proportional effect of age at diagnosis**
M1	Linear proportional effect of age
M3	Linear non-proportional effect of age
M4	Non-linear proportional effect of age
M5	Non-linear non-proportional effect of age
S2	**Simulation scenario 2, non-linear non-proportional effect of age, non-proportional effects of categorical stage and deprivation**
M2	Non-linear non-proportional effect of age, non-proportional effects of categorical stage and deprivation
M6	Non-linear non-proportional effect of age, non-proportional effect of categorical deprivation
M7	Non-linear non-proportional effect of age, non-proportional effect of categorical stage
M8	Linear proportional effect of age, categorical stage and deprivation
M9	Non-linear non-proportional effect of age, proportional effect of categorical stage, non-proportional effect of categorical deprivation
M10	Non-linear non-proportional effect of age, non-proportional effect of categorical stage, proportional effect of categorical deprivation

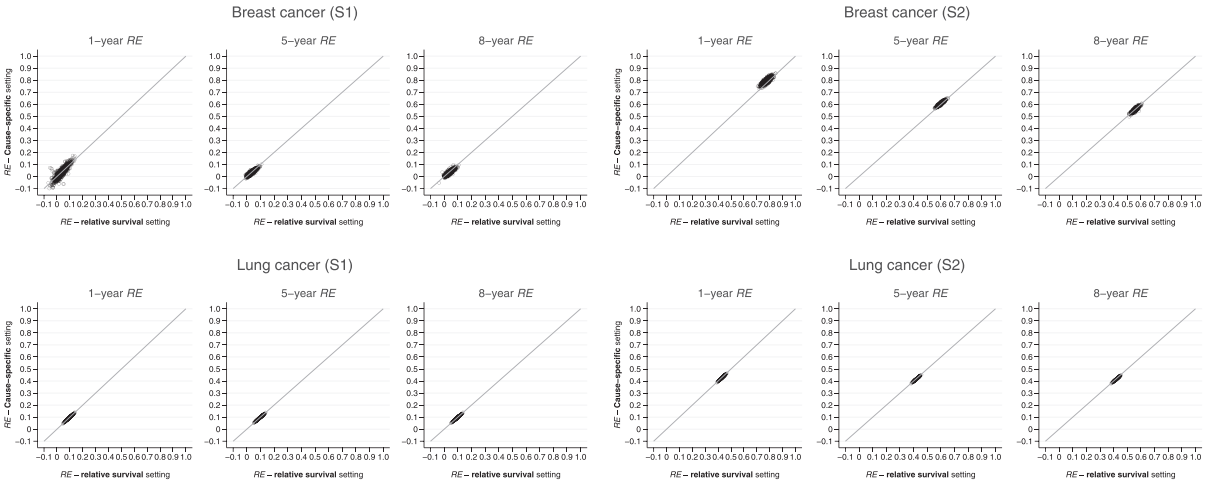


FIGURE 3 Comparison of RE obtained in cause-specific and relative survival settings, by cancer and simulation scenario

It is good practice to report the estimated number of cancer deaths, and their proportion among all deaths, as estimated by the sum of weights, so the interpretation of the outputs is given the required caution. Some degree of instability in the estimates of effects is indeed expected in excess hazard models where there is a low proportion of cancer deaths among all deaths.[17] REw is based on the excess hazard model and therefore suffers twice (through weighting and ranking of events) in such situations. In practice, we follow the recommendation from Sasieni that excess hazard model is best used when the proportion of death due to the disease of interest is between 30% and 90%.[20]

3.3 | REw—Properties

Mis-specifying the form of the effects of the main prognostic factors hardly affected REw. In simulation scenario S1 (simple linear proportional effect of age), the over-parameterisation of age in the modelling, by inclusion of non-linear and/or non-proportional effects of age (models M3-M5), did not alter REw: median REw, at 0.035 (breast) and 0.095 (lung) with model M1, increased to 0.037 to 0.043 (breast) and remained unchanged for lung (Figure 4).

The impact of stage, a strong predictor of survival, on REw is obvious when stage was omitted in the modelling (M6) while it was present in the simulation scenario (S2): median REw decreased from 0.600 (M2) to 0.213 (M6) for breast, and from 0.421 (M2) to 0.126 (M6) for lung (Figure 4). All other types of model mis-specification, such as omitting deprivation (M7), or omitting/including non-linearity or non-proportionality of age, deprivation or stage (M8-M10), did not have any strong impact on REw: for both breast and lung cancers, the largest differences in median REw occurred with under-parameterisation of stage, ie, lack of non-proportionality of the effect (M8, M9), and still showed a difference in median REw as small as 0.02 or less.

REw is robust to model mis-specification because the ranking of the individual hazards is unaffected by estimated changes in the strength of the effects only. M6, in which the effect of stage is ignored, shows greater impact on REw due to large changes in the ranking of observations.

The local REw was calculated using 20 events around each index event. This choice resulted in windows of varying lengths: stable at around 25 days all through the follow-up for breast cancer patients, whereas it started at less than 20 days for the first year of lung cancer follow-up, and then gradually increased to 450 days beyond 7 years. Indeed, over 75% of deaths occurred in the year following the lung cancer diagnosis, although it takes 5 years to observe 75% of deaths for breast cancer patients.

There was little variation between the 1000 local REw curves when simulation and estimation models coincided, ie, M1 in S1 and M2 in S2 (Figure 5). The general patterns of local REw seen in well-specified models were however preserved for mis-specified estimation models. In the simple scenario S1, local REw remained relatively constant with time since diagnosis for all models. For the more complex simulation scenario S2, the local REw curves decreased with time for all models. We further explored that decrease in local REw to understand what effect it reflected. We looked at simulated data following 2 additional scenarios, S3 and S4: S3 included linear proportional effects of age, and proportional effects of categorical deprivation and stage at diagnosis, while S4 included non-linear and non-proportional effects of age,

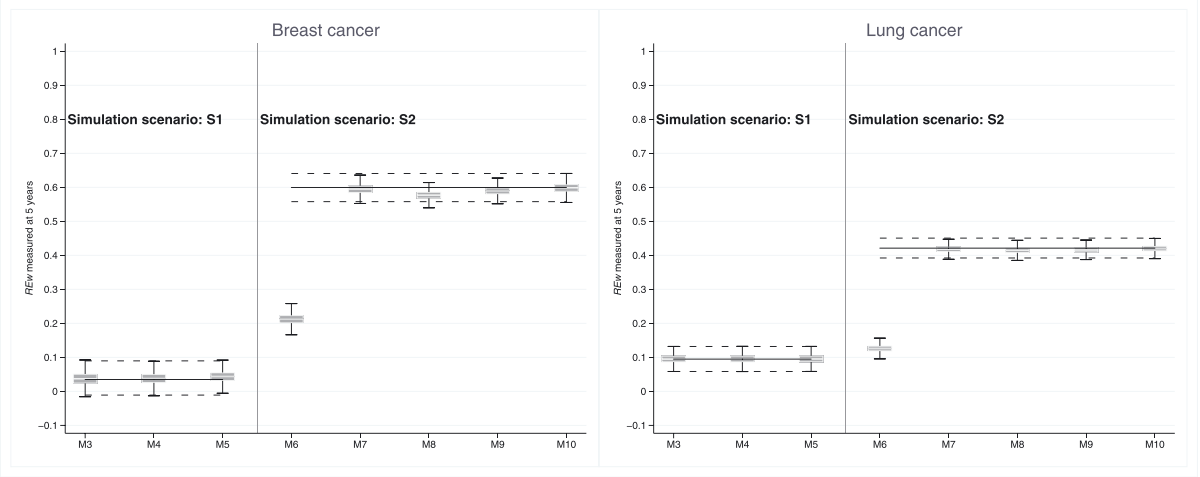


FIGURE 4 REw measured at 5 years, using different well-specified (M1, M2, plain lines) and mis-specified (M3-M10) models, by cancer and simulation scenario. S1: Simulation scenario 1, linear proportional effect of age at diagnosis; M1: Linear proportional effect of age (plain line across M3-M5); M3: Linear non-proportional effect of age; M4: Non-linear proportional effect of age; M5: Non-linear non-proportional effect of age; S2: Simulation scenario 2, non-linear non-proportional effect of age, non-proportional effects of categorical stage and deprivation (plain line across M6-M10); M2: Non-linear non-proportional effect of age, non-proportional effects of categorical stage and deprivation; M6: Non-linear non-proportional effect of age, non-proportional effect of categorical deprivation; M7: Non-linear non-proportional effect of age, non-proportional effect of categorical stage; M8: Linear proportional effect of age, categorical stage and deprivation; M9: Non-linear non-proportional effect of age, proportional effect of categorical stage, non-proportional effect of categorical deprivation; M10: Non-linear non-proportional effect of age, non-proportional effect of categorical stage, proportional effect of categorical deprivation

and non-proportional effects of categorical deprivation. While the local REw curves also decreased in simulation scenario S4, they remained constant in S3 indicating that non-proportional effects of age and other factors, rather than the adjustment for stage, triggered a decreasing local REw in S2.

The weighting initially proposed in Stare et al takes good account of random censoring through time. However, in the event that some cohorts of patients are all censored at a fixed date, such as due to administrative censoring when performing a complete study design, this weighting was not sufficient to cope for such large amount of censored information, often tied. We advise to break ties by simply adding or subtracting a small fraction of time to the survival times that tie. It will then prevent the spurious increasing REw emanating from large proportions of patients censored at similar times after diagnosis. Without this correction, local and overall REw will also converge to 1 from the time heavy censoring starts occurring. For users analysing a cohort study design, it is advised to measure the cumulative REw right before the administrative censoring happens.

4 | APPLICATION: COMPLEX MULTIVARIABLE MODELLING

Given the availability of potential predictors of cancer survival in England, we selected patients diagnosed with colon cancer in 2011 to 2013 ($n = 9300$) or non-small cell lung cancer in 2012 ($n = 5958$), with follow-up until the end of 2014. We selected a 25% random sample of patients with valid information on sex, age at diagnosis, deprivation, stage at diagnosis, major surgical treatment, and comorbidity (Charlson index, CCI) for all patients and additional information on performance status and route to diagnosis for lung cancer patients only (Table 1).

The initial parametric log-cumulative excess hazard models, stratified by sex, included age at diagnosis and deprivation, and expected hazards were provided by life tables defined by sex, single year of age, and deprivation. We aimed to measure the explained variation of the increasingly more complex models to reflect the explained variation of each factor successively added into the models.

The sum of the weights derived for the calculation of REw quantified the proportion of cancer deaths among all deaths. Of the 40.6% (42.0%) colon cancer patients dying through the follow-up, we estimated that 79.0% (83.4%) died

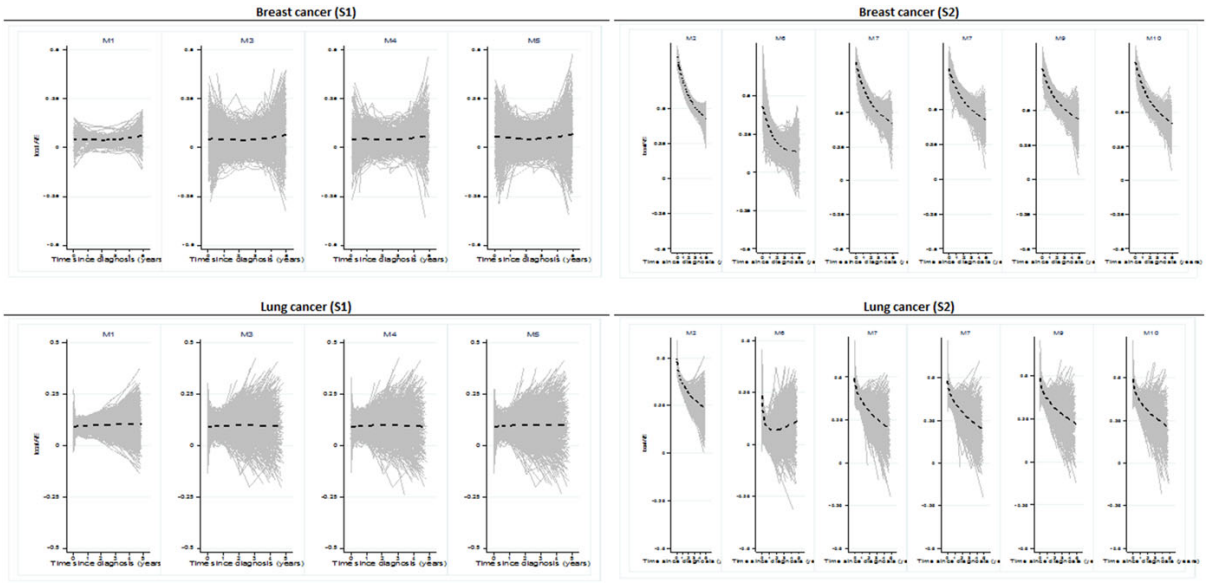


FIGURE 5 Local REw measured up to 5 years, for different well-specified (M1, M2) and mis-specified models (M3-M5 and M6-M10): breast and lung cancers. S1: Simulation scenario 1, linear proportional effect of age at diagnosis; M1: Linear proportional effect of age; M3: Linear non-proportional effect of age; M4: Non-linear proportional effect of age; M5: Non-linear non-proportional effect of age; S2: Simulation scenario 2, non-linear non-proportional effect of age, non-proportional effects of categorical stage and deprivation; M2: Non-linear non-proportional effect of age, non-proportional effects of categorical stage and deprivation; M6: Non-linear non-proportional effect of age, non-proportional effect of categorical deprivation; M7: Non-linear non-proportional effect of age, non-proportional effect of categorical stage; M8: Linear proportional effect of age, categorical stage and deprivation; M9: Non-linear non-proportional effect of age, proportional effect of categorical stage, non-proportional effect of categorical deprivation; M10: Non-linear non-proportional effect of age, non-proportional effect of categorical stage, proportional effect of categorical deprivation

of cancer in men (women); and of the 86.4% (81.3%) of dead men (women) lung cancer patients, 94.3% (95.0%) died of their cancer.

Table 2 shows that REw reached 0.22 (95%CI: 0.16–0.28) (REw = 0.26, 95%CI: 0.20–0.31) in men (women) with colon cancer, and 0.14 (95%CI: 0.11–0.17) (REw = 0.16, 95%CI: 0.12–0.19) in men (women) with lung cancer at 12 months after diagnosis, for models adjusted for age and deprivation only, ie, the baseline model. Full adjustment for all available covariables increased REw to 0.69 (95% CI: 0.67–0.70) (REw = 0.67, 95%CI: 0.66–0.69) in men (women) with colon cancer and 0.56 (95%CI: 0.54–0.58) (REw = 0.58, 95%CI: 0.56–0.60) in men (women) with lung cancer. Stage accounted for most of the increase in colon cancer, explaining 61.8% (53.5%) in men (women) of the explained variation of the full model, and increasing the baseline REw by over 150%. In lung cancer, performance status and stage showed the largest increase in REw, from the minimum initial model: around 200%, with an absolute change in REw of 0.29 (0.30) and 0.28 (0.29) in men (women) respectively; but in a full model, their respective shares represented 12.4% (10.8%) and 10.5% (7.4%) in men (women), suggesting correlation between variables such as treatment and stage, or emergency presentation and stage.

We then measured time-varying REw at 1 month and every 3 months following diagnosis, up to 3 years (Figures 6A and 7A). In colon cancer patients, there is a clear distinction between models that do and do not contain stage at diagnosis. In models excluding stage, REw(t) was stable from 12 months after a sharp decrease in the first 6 months and then slight decrease until the 12^{th} month (Figure 6A). The local REw showed evidence that at 2 years after diagnosis, models that contained the surgical treatment variable, without stage at diagnosis, displayed an increased local REw (Figure 6B).

In lung cancer models, the time-varying REw increased from the baseline age and deprivation model with any additional variable: REw(t) was stable after a slight decrease until 3 months, mostly in women, and in models adjusted for emergency presentation (from over 0.5 in women to less than 0.4, Figure 7A). Additionally, the curves reflecting presence of stage and performance status reflect perfectly the large contribution of performance status at the start of the follow-up, and the constant contribution of stage. The patterns of the local REw curves are suggestive of a late treatment

TABLE 1 Number and proportion of patients, by stage at diagnosis and each of the main explanatory factors: Lung cancer patients diagnosed in 2012, colon cancer patients diagnosed in 2011 to 2013 in England

	Non-Small Cell Lung Cancer																			
	Men										**Women**									
	Stage I		**Stage II**		**Stage III**		**Stage IV**		**Total**		**Stage I**		**Stage II**		**Stage III**		**Stage IV**		**Total**	
Age at diagnosis																				
Mean (sd)	72.8	(10)	72.9	(10.2)	72.1	(10.06)	72.3	(10.4)			72.6	(10.7)	72.6	(11)	72.1	(10.5)	72.2	(11.1)		
Treatment	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
No major surgical treatment	223	44.2	149	52.3	753	89.1	1658	99.0	2783	84.1	214	44.7	123	52.1	528	89.0	1327	98.9	2192	82.7
Major surgical treatment	281	55.8	136	47.7	92	10.9	16	1.0	525	15.9	265	55.3	113	47.9	65	11.0	15	1.1	458	17.3
Emergency presentation																				
No	425	84.3	237	83.2	657	77.8	1016	60.7	2335	70.6	367	76.6	186	78.8	452	76.2	767	57.2	1772	66.9
Yes	79	15.7	48	16.8	188	22.2	658	39.3	973	29.4	112	23.4	50	21.2	141	23.8	575	42.8	878	33.1
Performance status																				
High—0	168	33.3	87	30.5	201	23.8	230	13.7	686	20.7	138	28.8	66	28.0	122	20.6	177	13.2	503	19.0
1	186	36.9	114	40.0	321	38.0	521	31.1	1142	34.5	171	35.7	86	36.4	220	37.1	435	32.4	912	34.4
2	83	16.5	47	16.5	147	17.4	388	23.2	665	20.1	77	16.1	48	20.3	108	18.2	268	20.0	501	18.9
3	51	10.1	27	9.5	121	14.3	370	22.1	569	17.2	63	13.2	29	12.3	115	19.4	306	22.8	513	19.4
4	7	1.4	3	1.1	36	4.3	126	7.5	172	5.2	12	2.5	3	1.3	20	3.4	124	9.2	159	6.0
Low—5	9	1.8	7	2.5	19	2.2	39	2.3	74	2.2	18	3.8	4	1.7	8	1.3	32	2.4	62	2.3
Deprivation quintile																				
Least deprived	81	16.1	35	12.3	107	12.7	241	14.4	464	14.0	54	11.3	32	13.6	74	12.5	165	12.3	325	12.3
2	84	16.7	51	17.9	139	16.4	269	16.1	543	16.4	85	17.7	39	16.5	93	15.7	194	14.5	411	15.5
3	93	18.5	56	19.6	161	19.1	355	21.2	665	20.1	83	17.3	50	21.2	130	21.9	264	19.7	527	19.9
4	100	19.8	70	24.6	224	26.5	365	21.8	759	22.9	119	24.8	50	21.2	141	23.8	367	27.3	677	25.5
Most deprived	146	29.0	73	25.6	214	25.3	444	26.5	877	26.5	138	28.8	65	27.5	155	26.1	352	26.2	710	26.8
Charlson comorbidity score																				
None—0	256	50.8	165	57.9	524	62.0	1054	63.0	1999	60.4	251	52.4	136	57.6	373	62.9	917	68.3	1677	63.3
1	113	22.4	52	18.2	141	16.7	295	17.6	601	18.2	127	26.5	55	23.3	107	18.0	218	16.2	507	19.1
2	54	10.7	37	13.0	83	9.8	169	10.1	343	10.4	47	9.8	28	11.9	56	9.4	98	7.3	229	8.6
>2	81	16.1	31	11.0	97	11.3	156	9.4	365	11.0	54	11.2	17	7.2	57	9.6	109	7.9	237	8.9
Total	504	100.0	285	100.0	845	100.0	1674	100.0	3308	100.0	479	100.0	236	100.0	593	100.0	1342	100.0	2650	100.0

	Colon cancer																			
	Men										**Women**									
	Stage I		**Stage II**		**Stage III**		**Stage IV**		**Total**		**Stage I**		**Stage II**		**Stage III**		**Stage IV**		**Total**	
Age at diagnosis																				
Mean (sd)	70.9	(11.2)	72.5	(11.6)	69.8	(12.6)	72.0	(12)			70.5	(13.3)	73.8	(12)	72.1	(12.7)	71.9	(13.7)		
Treatment	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
No major surgical treatment	31	4.7	41	2.8	59	4.5	393	25.4	524	10.6	27	4.9	49	3.7	68	6.0	404	30.1	548	12.6
Major emergency surgery	33	5.0	264	18.2	268	20.6	258	16.7	823	16.6	35	6.3	258	19.7	283	24.8	241	17.9	817	18.8
Major elective surgery	421	64.4	1028	71.0	845	64.9	321	20.8	2615	52.8	377	67.8	914	69.7	680	59.7	256	19.1	2227	51.2
Minor surgery	169	25.8	115	7.9	131	10.1	573	37.1	988	20.0	117	21.0	91	6.9	108	9.5	442	32.9	758	17.4
Deprivation quintile																				
Least deprived	140	21.4	318	22.0	290	22.3	331	21.4	1079	21.8	131	23.6	304	23.2	291	25.5	262	19.5	988	22.7
2	154	23.5	325	22.4	307	23.6	343	22.2	1129	22.8	118	21.2	290	22.1	247	21.7	293	21.8	948	21.8
3	150	22.9	318	22.0	266	20.4	329	21.3	1063	21.5	121	21.8	270	20.6	215	18.9	280	20.8	886	20.4
4	117	17.9	269	18.6	251	19.3	321	20.8	958	19.4	114	20.5	238	18.1	205	18.0	301	22.4	858	19.7
Most deprived	93	14.2	218	15.1	189	14.5	221	14.3	721	14.6	72	12.9	210	16.0	181	15.9	207	15.4	670	15.4
Charlson comorbidity score																				
None - 0	477	72.9	1061	73.3	1015	77.9	1146	74.2	3699	74.7	440	79.1	1037	79.0	923	81.0	1076	80.1	3476	79.9
1	78	11.9	194	13.4	122	9.4	179	11.6	573	11.6	65	11.7	142	10.8	106	9.3	139	10.3	452	10.4
2	49	7.5	108	7.5	91	7.0	116	7.5	364	7.4	31	5.6	78	5.9	62	5.4	66	4.9	237	5.4
>2	50	7.7	85	5.8	75	5.8	104	6.8	314	6.3	20	3.7	55	4.3	48	4.3	62	4.6	185	4.3
Total	654	100.0	1448	100.0	1303	100.0	1545	100.0	4950	100.0	556	100.0	1312	100.0	1139	100.0	1343	100.0	4350	100.0

TABLE 2 Multivariable model: Explained variation (*REw*) measured at 12 months after diagnosis, for overall models and individual variables

				Change[a] in REw			
				Inclusion[b]		Exclusion[b]	
		REw at 12 months (95% CI)		Difference in REw	Proportion of Initial Model (%)	Difference in REw	Proportion of Full Model (%)
Colon cancer							
Men							
Initial model:	*Age, deprivation*	*0.221*	*(0.160; 0.282)*				
	Age, deprivation, stage	0.671	(0.657; 0.686)	0.450	203.5	0.427	61.8
	Age, deprivation, treatment[c]	0.251	(0.195; 0.307)	0.030	13.5	0.016	2.4
	Age, deprivation, Charlson Comorbidty index (CCI)	0.232	(0.171; 0.292)	0.010	4.7	0.003	0.4
Full model:	*Age, deprivation, stage, treatment, CCI*	*0.690*	*(0.676; 0.704)*				
Women							
Initial model:	*Age, deprivation*	*0.256*	*(0.198; 0.314)*				
	Age, deprivation, stage	0.660	(0.644; 0.675)	0.403	157.5	0.359	53.5
	Age, deprivation, treatment[c]	0.290	(0.241; 0.340)	0.034	13.4	0.010	1.4
	Age, deprivation, Charlson Comorbidity index (CCI)	0.271	(0.214; 0.329)	0.015	6.0	0.002	0.3
Full model:	*Age, deprivation, stage, treatment, CCI*	*0.671*	*(0.656; 0.686)*				
Non-small cell lung cancer							
Men							
Initial model:	*Age, deprivation*	*0.141*	*(0.112; 0.171)*				
	Age, deprivation, stage	0.422	(0.403; 0.441)	0.280	198.5	0.058	10.5
	Age, deprivation, treatment[d]	0.257	(0.235; 0.280)	0.116	81.9	0.003	0.6
	Age, deprivation, Charlson Comorbidity index (CCI)	0.141	(0.111; 0.170)	−0.001	−0.5	0.000	0.1
	Age, deprivation, performance status (PS)	0.434	(0.409; 0.459)	0.293	207.3	0.069	12.4
	Age, deprivation, presentation (EP vs non-EP)	0.325	(0.295; 0.354)	0.183	129.8	0.013	2.4
Full model:	*Age, deprivation, stage, treatment, CCI, PS, presentation*	*0.558*	*(0.539; 0.576)*				
Women							
Initial model:	*Age, deprivation*	*0.155*	*(0.120; 0.191)*				
	Age, deprivation, stage	0.442	(0.421; 0.463)	0.287	185.1	0.043	7.4
	Age, deprivation, treatment[d]	0.299	(0.274; 0.324)	0.144	92.8	0.006	1.0
	Age, deprivation, Charlson Comorbidity index (CCI)	0.157	(0.121; 0.192)	0.002	1.0	−0.001	−0.2
	Age, deprivation, performance status (PS)	0.455	(0.427; 0.484)	0.300	193.6	0.063	10.8
	Age, deprivation, presentation (EP vs non-EP)	0.352	(0.318; 0.387)	0.197	127.2	0.020	3.4
Full model:	*Age, deprivation, stage, treatment, CCI, PS, presentation*	*0.584*	*(0.564;0.604)*				

effect: generally decreasing over time for models including emergency presentation or performance status but increasing for the model including surgical treatment information (Figure 7B).

By definition of the local REw, the shapes of the smoothed curves are only slightly influenced by the number of events included in the windows around each index event. For both colon and lung cancers, including 10 events on either side of the index event resulted in windows of times varying between a day and 50 days for lung cancer or between a day and over 75 days for colon cancer in the 3 years of follow-up. The degree of smoothing will also likely impact the shape of the local REw curves. Furthermore, the cumulative nature of the overall and time-varying REw means that they are likely impacted by the high proportions of death happening at the beginning of the follow up: 50% of all deaths occurred by the 3rd and 9th months of follow-up in lung and colon cancer, respectively, explaining why REw(t) was mostly flat beyond these times.

5 | DISCUSSION

We presented here an adaptation of the RE measure for event history data to excess hazard modelling. We offer a new tool to quantify the variation in disease-specific outcome explained by the available predictive factors. In this context, REw can be measured at given time points following diagnosis and plotted as a function of time. Additional exploratory insight is provided by a "local REw", calculated using a window of events around each event time. That function of time can be very unstable, and the smoothed curve is useful to look at the general trend in the variation in RE by the model. Although dependent on death patterns, these time-varying versions of REw help understand better when specific factors have strongest impact on survival.

The differences between local REw and REw(t) curves can be seen similarly to the differences between hazard and cumulative hazard curves. The cumulative hazard curve is a cumulative measure, whereby hardly affected by local effects seen in the instantaneous hazard curve. REw(t) is the cumulative REw, heavily impacted by the first few months following the diagnosis, where most cancer-related deaths occur. If one is interested in changes in explained variation due to, say late treatment effects or changes in the composition of the cohort of patients (younger ages, fewer late stage patients, fitter patients...), the local REw will provide such information. Furthermore, in the context of dynamic data and dynamic models, local REw will be providing the necessary time-varying measure of explained variation.

Furthermore, local REw and REw(t) are informative for comparison between studies, or when varying follow-up times are available, because the overall measure REw will vary with the available follow-up.[21]

Further research in the number of events to include in the calculation of the local REw show very little variation in the smoothed functions. Only the heights of the spikes seen in the un-smoothed local REw curves are affected, and hence the tail of the smoothed curve, where the number of events is more scarce. We advise researchers to use 20 events, as a default size, and depending on the cancer lethality, check for the impact of using much smaller (say 4–10 events) or much larger (30–40 events) number of events. The local REw curve will, to some extent, depend on the number of events as well as the amount of smoothing applied.

The weighting system proposed here for the relative survival setting keeps the simplicity and the intuition of the original RE used in the overall and cause-specific settings. It also retains the original RE measure's properties such as model-free interpretation. Furthermore, the weighted measure REw in the relative survival setting is equivalent to RE in the cause-specific setting.

Multidimensional models defined on the log or log-cumulative hazard scales can now be routinely used to estimate excess hazard from cancer.[3,11] These models often include complex non-linear and non-proportional effects of a variety of factors that may impact levels of survival. Therefore, the regression coefficients are not straightforward to interpret,

Notes to Table 2

Modelled effects: Age: non-linear and non-proportional, Deprivation: categorical, non-proportional, Stage: categorical, non-proportional, Treatment: categorical, non-proportional, CCI: linear, non-proportional, Performance status: categorical, Presentation binary: emergency presentation (EP) versus non-emergency.

[a]Change is measured as the arithmetic difference between the initial (inclusion) or full (exclusion) model REw and the model that includes the specific variable. That difference is expressed as a proportion of the initial (inclusion) or full (exclusion) model.

[b]Inclusion: change in REw with the addition of the index variable to a model including age and deprivation. Exclusion: change in Rew with the removal of the index variable from the full model.

[c]The variable "treatment" represents major surgical resection.

[d]The variable "treatment" represents both treatment and the route to diagnosis: 1—no treatment, 2—emergency major surgery, 3—elective major surgery, 4—minor surgery.

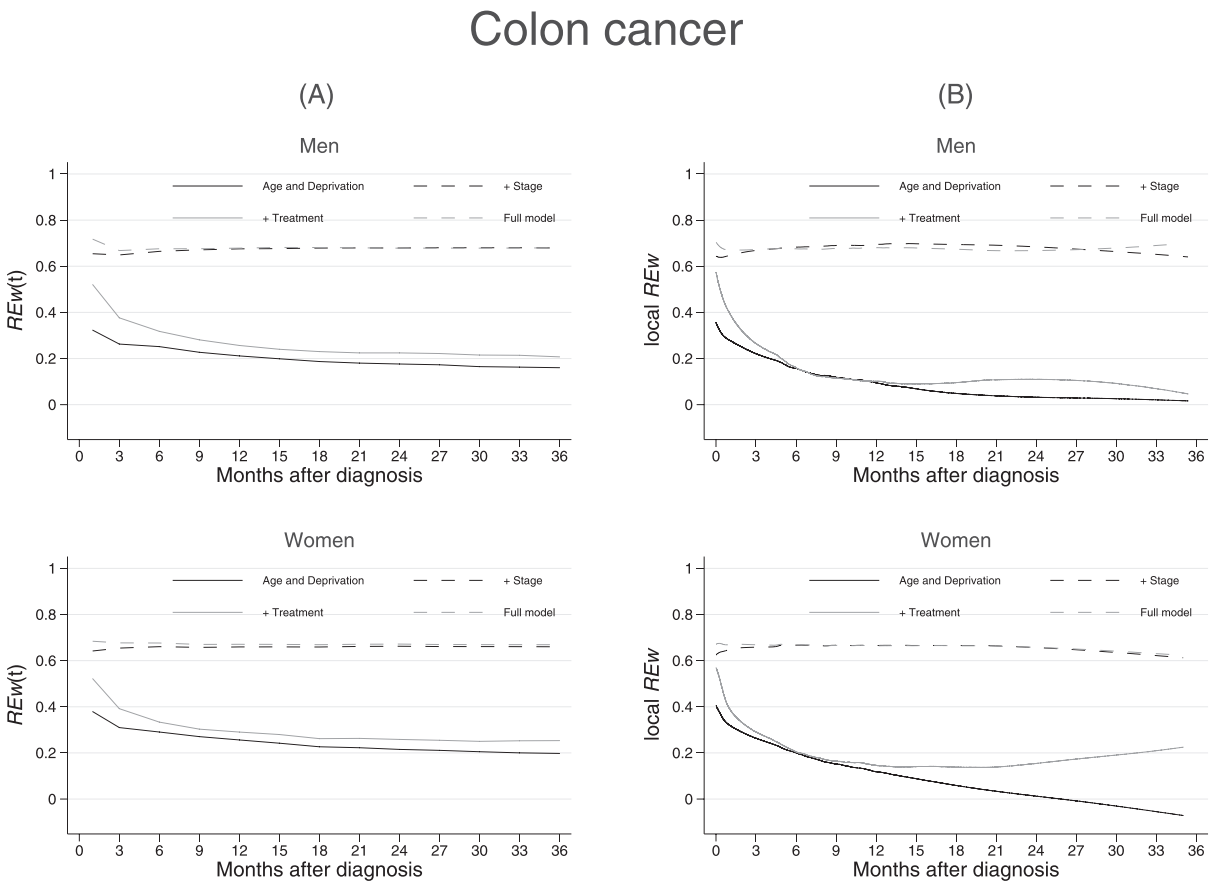


FIGURE 6 Multivariable models: (A) explained variation measured at 1 month and every 3 months after diagnosis, (B) smoothed local RE up to 3 years after diagnosis, for models adjusted for the effects of age and deprivation, and stage, and treatment. Colon cancer patients diagnosed in 2011 to 2013, selected for their valid stage at diagnosis: 4950 men and 4350 women the curve for comorbidity is not presented here as it is undistinguishable from the age and deprivation model

Notes: (1) RE(t) and local RE can have values between −1 and +1 (2) Cumulative RE, RE(t) is calculated at month 1, 3, 6...36 after diagnosis (3) Local RE is calculated using information from 10 events on either side of the index event. The smoothed (lowess with mean smoother) curve is presented here

and strong predictors are often hard to pin down. We propose to look at differences in REw between models to quantify the proportion of variation explained by a given factor. Our illustration shows that for lung cancer patients, performance status explained the largest amount of variation in survival between patients, particularly in the early months following diagnosis. Performance status, although well-known and discussed in Multi-Disciplinary Team meetings, is rarely accounted for in epidemiology, mainly because of its unavailability in the routine cancer registration datasets. Such high explanatory power for that variable could trigger its availability at least in specialised cancer registry datasets.

Low proportions of explained variation for single covariable in the full model, whereas each additional variable adds, individually, a lot to the explained variation of the baseline model, indicate high correlation between factors. It could reflect a high adherence to guidelines such that whole groups of patients got administered the same treatment, or were diagnosed via a given ideal route.

Despite the measure being dependent on the excess hazard model through the weighting and through the ranking of observations, REw proved a great stability to model specification. REw was largely insensitive to over-parameterisation or under-parameterisation of the variables used in the simulation model. Non-linear or non-proportional effects, although they may reflect better the reality of the estimated disease-specific survival, will not impact dramatically the order at which patients will experience the event of interest.

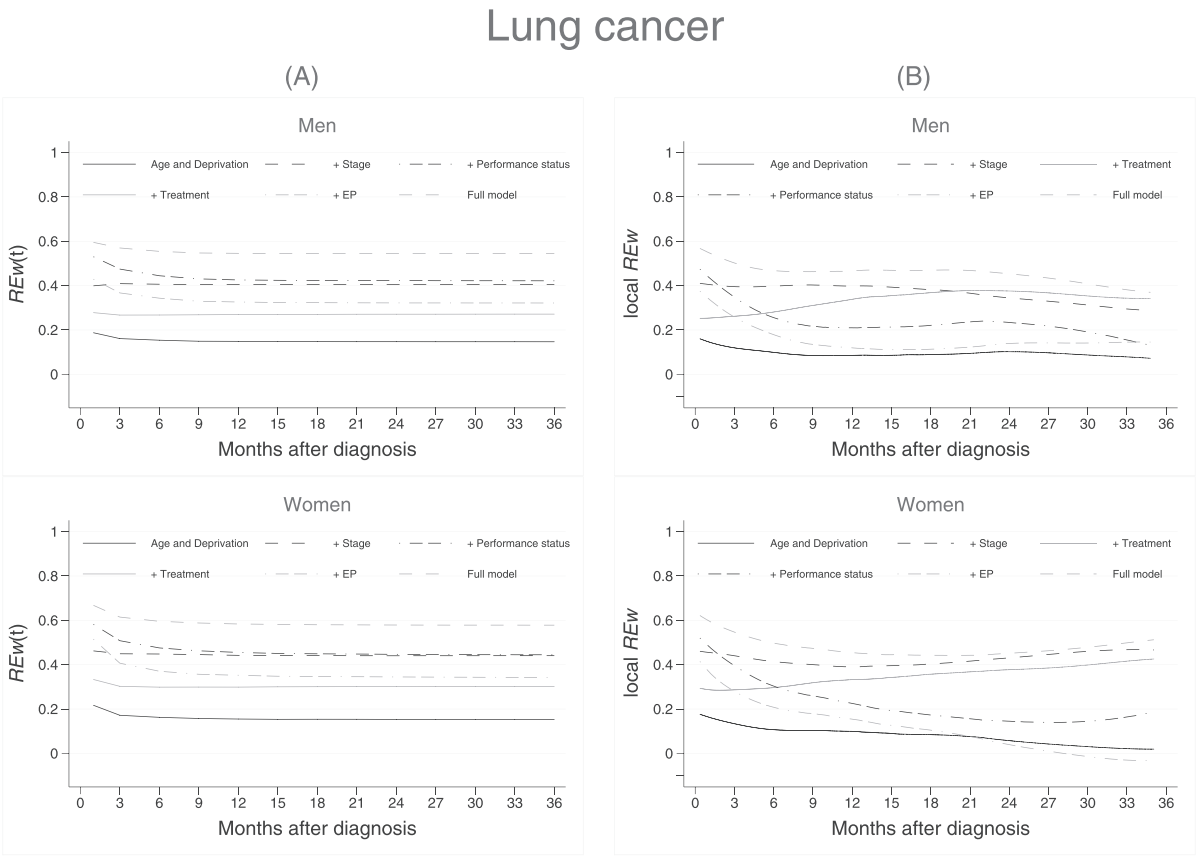


FIGURE 7 Multivariable models: (A) explained variation measured at 1 month and every 3 months after diagnosis, (B) smoothed local RE up to 3 years after diagnosis, for models adjusted for the effects of age and deprivation, and stage, treatment, performance status, and emergency presentation. Non-small cell lung cancer patients diagnosed in 2012, selected for their valid stage and performance status at diagnosis: 3308 men and 2650 women. The curve for comorbidity is not presented here as it is undistinguishable from the age and deprivation model
Notes: (1) RE(t) and local RE can have values between −1 and +1 (2) Cumulative RE, RE(t) is calculated at month 1, 3, 6...36 after diagnosis (3) Local RE is calculated using information from 10 events on either side of the index event. The smoothed (lowess with mean smoother) curve is presented here

REw, like RE, is not exact. Small sample sizes or low number of deaths due to the disease of interest will increase variability around the estimated REw. Therefore, we advise users to report the variance or confidence interval obtained around the estimated REw. Similar to RE, REw estimates may be biased for a factor with a small effect.[13] However, the bias will become negligible as the sample size increases.

Further developments will include testing the REw on dynamic models that include time-varying variables,[22] and in hierarchical models.[2] The variation explained by these models may be greater, because they allow the effect of time-varying variables to be modelled and, hence, measures of prognostic factors that are updated over time since the cancer diagnosis.

ACKNOWLEDGMENTS

We are grateful to CRUK for funding this research.

ORCID

Camille Maringe http://orcid.org/0000-0002-8739-9565
Maja Pohar Perme http://orcid.org/0000-0002-3412-2642

REFERENCES

1. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *The Stata Journal*. 2009;9:265-290.
2. Charvat H, Remontet L, Bossard N, et al. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Stat Med*. 2016;35(18):3066-3084.
3. Remontet L, Bossard N, Belot A, et al. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Stat Med*. 2007;26(10):2214-2228.
4. Giorgi R, Abrahamowicz M, Quantin C, et al. A relative survival regression model using B-spline functions to model non-proportional hazards. *Stat Med*. 2003;22(17):2767-2784.
5. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med*. 1999;18(17–18):2529-2545.
6. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics*. 2005;61(1):92-105.
7. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med*. 2004;23(13):2109-2123.
8. Schemper M, Stare J. Explained variation in survival analysis. *Stat Med*. 1996;15(19):1999-2012.
9. Cox DR. Regression models and life-tables. *J R Stat Soc B Methodol*. 1972;34(2):187-220.
10. Esteve J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. *Stat Med*. 1990;9(5):529-538.
11. Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. *Stat Med*. 2007;26(30):5486-5498.
12. Danieli C, Remontet L, Bossard N, et al. Estimating net survival: the importance of allowing for informative censoring. *Stat Med*. 2012;31(8):775-786.
13. Stare J, Perme MP, Henderson R. A measure of explained variation for event history data. *Biometrics*. 2011;67(3):750-759.
14. Harrell FE Jr, Califf RM, Pryor DB, et al. Evaluating the yield of medical tests. *JAMA*. 1982;247(18):2543-2546.
15. Schemper M, Henderson R. Predictive accuracy and explained variation in Cox regression. *Biometrics*. 2000;56(1):249-255.
16. Pohar Perme M, Stare J, Esteve J. On estimation in relative survival. *Biometrics*. 2012;68(1):113-120.
17. Pohar Perme M, Henderson R, Stare J. An approach to estimation in relative survival regression. *Biostatistics*. 2009;10(1):136-146.
18. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Stat Med*. 2013;32(23):4118-4134.
19. Crowther MJ, Lambert PC. Simulating complex survival data. *The Stata Journal*. 2012;12(4):674-687.
20. Sasieni PD. Proportional excess hazard. *Biometrika*. 1996;83(1):127-141.
21. Kejzar N, Maucort-Boulch D, Stare J. A note on bias of measures of explained variation for survival data. *Stat Med*. 2016;35(6):877-882.
22. Mauguen A, Rachet B, Mathoulin-Pelissier S, et al. Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models. *Stat Med*. 2013;32(30):5366-5380.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Maringe C, Pohar Perme M, Stare J, Rachet B. Explained variation of excess hazard models. *Statistics in Medicine*. 2018;1–17. https://doi.org/10.1002/sim.7645

2.6.2 The Brier score

In this section, we lay out the theory and initial ideas for the adaptation of the Brier score to the relative survival data setting. The suggestion stems from the prediction error introduced by Schoop et al. [137] and presented in equation 2.4 in section (2.3.1). We define $Z_i^* \subset Z_i$ a subset of the variables defining the life tables.

In the relative survival data setting,

1 The prediction $\pi^1(t|Z_i)$ would be estimated from an excess hazard model as the crude probability of cancer death, CPD.

2 $\mathbb{1}_{(\tilde{T}_i \leq t)}$ would need to be used in lieu of $\mathbb{1}_{(\tilde{T}_i \leq t, \delta_i^1 = 1)}$ given we do not know what cause contributed to observing an event.

We therefore need to add to the informative censoring weights w^1 a component that reflects the mixtures of deaths included in $\mathbb{1}_{(\tilde{T}_i \leq t)}$. Similarly to what was proposed for RE, the weights could be derived from the estimated individual excess hazards, such that it is the ratio of the excess hazard over the overall hazard, as estimated by the model and the life tables (population) hazards: $w^2(t, Z_i) = \frac{\lambda_{E,i}(t,Z_i)}{\lambda_{E,i}(t,Z_i)+\lambda_{P,i}(t,Z_i^*)}$.

Therefore we would have the following time-varying prediction error in the context of excess hazard modelling:

$$PE(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{1}_{(\tilde{T}_i \leq t, \delta_i = 1)} - C\hat{P}D_i^1(t|Z_i)\right]^2 * w^1(t, \tilde{T}_i, \delta_i, \hat{G}(t), Z_i) * w^2(t, Z_i). \quad (2.5)$$

2.6.3 The ROC curve

Lorent et al. [156] propose an adaptation of the time-dependent ROC curve for censored survival data to the net survival context. The measures of sensitivity and specificity are modified such that the estimation of the joint distribution of survival and the marker is replaced by an estimation of the joint distribution of net survival and the marker. The estimation method proposed is that of nearest neighbour estimation, relying on patients with similar marker values to inform on the survival probabilities.

Although this method is in-keeping with the proposals reviewed briefly here, it could be argued that the choice of marker could be improved. It is understood that the authors investigate the predictive capacity of a marker, defined broadly for overall survival, to predict disease-specific mortality. This could be replaced with a marker specifically defined for prediction of disease mortality, such as derived from an excess hazard model.

2.6.4 Sensitivity and Specificity measures

Similarly to section 2.6.2, we look here at how measures of sensitivity and specificity could be estimated in the relative survival data setting. We start by looking at the original description of the measures, in the binary response setting.

Classic binary response setting

The true positive rate or sensitivity is defined as $Se = \frac{Number\ true\ positive}{Number\ disease\ positive}$ and the true negative rate or specificity is defined as $Sp = \frac{Number\ true\ negative}{Number\ disease\ negative}$.

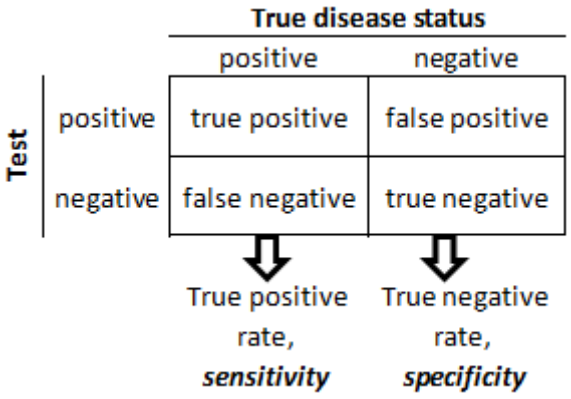


Figure 2.5: Sensitivity and specificity: binary outcome

Time-to-event setting

In the time-to-event setting, one repeats such classification, for given times t after diagnosis, and we compare the patients still at risk of the event, and those not at risk anymore. If there is censoring, then some patients may truly 'not be at risk' of the event anymore at time t, although they have not yet experienced the event itself prior to time t, due to censoring.

In this case, for some of these subjects, there is a mismatch for some patients between what the test results (low score – low probability to have experienced the event already) and the reality (not at risk). When this happens, Heagerty et al. [148] offer estimators that rely on non-parametric nearest neighbour estimation of the bivariate distribution of the time-to-event and marker processes, needed for the estimation of sensitivity and specificity.

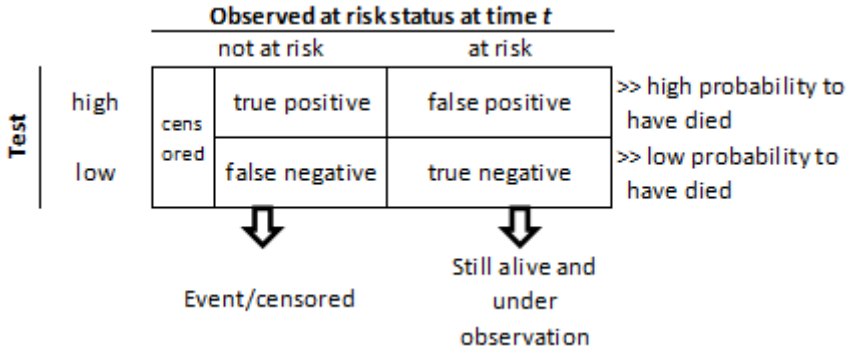


Figure 2.6: Sensitivity and specificity: time-to-event outcome

Relative survival data setting

In this setting, the group of patients observed not to be at risk of the event at time t is composed of: (i) Patients who experienced the event of interest prior to time t; (ii) Patients who were censored prior to time t; (iii) Patients who died of causes other than the cause of interest prior to time t.

In excess hazard models, we model the hazard of dying from the cause of interest, say cancer. Hence, a given patient's test marker may be high, if and only if the patient had a high probability of dying from cancer between diagnosis and time t. Therefore patients who died of other causes prior to time t may have a low test marker. Some patients may be at high risk of both types of death (cancer and other causes) and will therefore have a high test result, and not be at risk anymore.

I propose to weight each observations in the 'not at risk' group of patients, such that their contribution is not full but represent the probability that their observed event is the event of interest (weights proposed for RE). In some ways, we simply modify the number of observed events, to account for competing risks of death. If we denote by $w_i = \frac{\lambda_{E,i}}{\lambda_{E,i}+\lambda_{P,i}}$ the probability that the event observed in patient i is the event of interest, we define sensitivity as: $Se_w = \frac{\sum_{i=1}^{N} w_i * \mathbb{1}_{T_i<t, M_i>c}}{\sum_{i-1}^{N} w_i * \mathbb{1}_{T_i<t}}$ and I do not think we would need to modify the measure of specificity. In the case of additional censoring, specific further corrections will need to apply too.

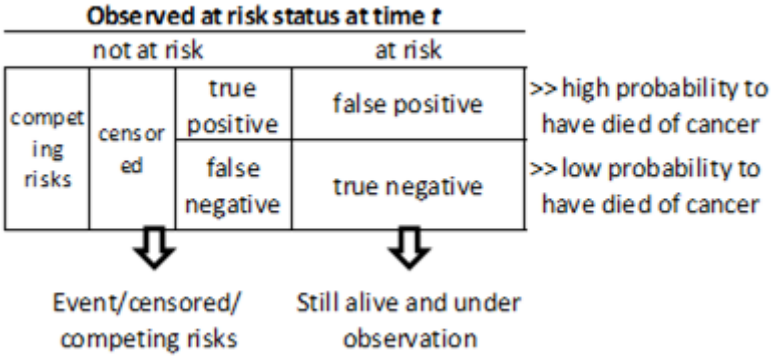


Figure 2.7: Sensitivity and specificity: time-to-event outcome, relative survival data setting

2.7 Conclusion

All measures covered in this Chapter aim at evaluating the prognostic characteristics of statistical models. They all carry complementary specific properties, akin to the different characteristics of explained variation, explained randomness and predictive accuracy, discussed in Choodari-Oskooei. [140, 141] Although measures of calibration and discrimination are typically useful for statistical models aimed at producing individual predictions, they could still be useful for population-based predictions. Calibration, sensitivity and specificity can be assessed for groups of patients, sharing specific characteristics. These measures would be relevant when measured in the sample of patients that trained the models, as well as on new cohorts of patients for whom survival is predicted. In a similar way, measures of predictive accuracy [140, 141] can be used for comparing predicted and observed outcomes in groups of patients defined by similar characteristics.

No specific measure had yet been developed or adapted to the context of the relative survival data setting. Of all measures presented in this Chapter, REw is the first measure that is readily accessible for use in the relative survival data setting. Although based on individual rankings at each time of event, this measure reflects the accuracy of model-based prediction at cohort level, or for a group of patients we are interested in.

Explained randomness measures are based on entropy, as is the Akaike Information Criteria (AIC). They are an indication of how precisely a model reproduces the patterns observed. In the next chapter we will use AIC to aid model-selection for the prediction of survival.

2.8 Further contribution to the topic

Stata program for cross-validated AUC

Luque-Fernandez MA, Redondo-Sánchez D, *Maringe C*. cvauroc: Command to compute cross-validated area under the curve for ROC analysis after predictive modeling for binary outcomes. The Stata Journal. 2019; 19(3):615-25

Chapter 3

Population-based predictions of cancer survival

3.1 Introduction

In this Chapter, we aim to select models for the prediction of cancer survival. The work presented here builds on previous Chapters. Firstly, it shows how the algorithms presented in Chapter 1 are adapted for the specific purpose of prediction. We also measure the variation in outcomes explained by the selected model(s) using REw, developed and defined in Chapter 2.

As described in Burnham and Anderson [157] (page 284), there are usually three different ways to pursue model selection:

1. Tests of hypotheses
2. Optimisation of small selection criteria
3. Ad-hoc methods

Chapter 1 describe in details two classes of algorithms based on hypothesis testing using likelihood ratio tests. These algorithms are adapted to excess hazard models. In this Chapter, we aim to investigate how model selection based on the optimisation of Information Criteria (IC) provides a relevant framework when prediction of survival is intended.

As highlighted in a commentary by A.E. Raftery, [158] the tests of hypotheses such as likelihood ratio tests are ill-defined for model selection, especially in large samples. Tests will detect discrepancies between the model and the data, rather than express how close a

model is to the data: the larger the sample, the easier it is to detect (small) discrepancies. In the context of the analysis of population-based cancer survival, datasets may be large, and as seen in Chapter 1, most selected models tend to be complex. This is one reason why we now turn to the use of information criteria, and specifically the Akaike Information Criterion (AIC) [159] and the Bayesian Information Criterion (BIC). [160]

We start by describing the Kullback-Leibler distance, a measure that compares the distance between two probability distributions. We then define the AIC as an estimator of the relative Kullback-Leibler distance, and the BIC. We present how we adapted the Royston and Sauerbrei algorithm to model-selection using information criteria and how we make multi-model inference from the selected models.

3.2 Information criteria

The Kullback-Leibler (K-L) distance is a directed measure of information, or difference between two distributions. We assume that we are trying to approximate the true distribution f with a probability distribution function g. The K-L distance provides an idea of how much information is lost when g is used instead of f. Its calculation relies on the fact that f is fully specified and the parameters θ of the distribution function g are fully known. As such, it cannot be used if one tries to approximate an unknown distribution. It is a directed distance, meaning that the information we lose when we use g instead of f is not the same as the information lost when we approximate g by f: this is seen in the asymmetry of the formula below for the K-L distance:

$$I(f,g) = \int f(X) log\left(\frac{f(X)}{g(X|\theta)}\right) dx \tag{3.1}$$

One can rearrange the above formula to

$$I(f,g) = \int f(X) log\left(f(X)\right) dx - \int f(X) log\left(g(X|\theta)\right) dx \tag{3.2}$$

and

$$I(f,g) = \mathbb{E}_f\left[log(f(X))\right] - \mathbb{E}_f\left[log(g(X|\theta))\right] \tag{3.3}$$

$\mathbb{E}_f\left[log(f(X))\right]$ is only dependent on the true probability distribution f that we are trying to approximate. Whatever the function g, $\mathbb{E}_f\left[log(f(X))\right]$ is constant, and the value of $I(f,g)$ will vary according to values of $-\mathbb{E}_f\left[log(g(X|\theta))\right]$. This part of the equation is the relative K-L distance between f and g, see Burnham and Anderson [157] (page 58). Some considerations to keep in mind, in relation to the relative distance:

(a) It lacks a true zero: the minimum value for the relative distance is obtained when $-\mathbb{E}_f\left[log(g(X|\theta))\right]$ reaches $\mathbb{E}_f\left[log(f(X))\right]$.

(b) Whatever the sample size N of X, a given difference between two relative distances will have the same meaning.

The vector of parameters θ_0 that minimise the K-L distance between f and g, depend on f, g and the sample of data available, X.

3.2.1 AIC

In the discussion above on the Kullback-Leibler distance and its simplification to the relative K-L distance, we assume known both f and g and their parameters. There is a true value of parameters θ_0 for model g that minimise the relative distance, for a given g. In reality, we do not know what the parameters θ_0 of g are, and these must be estimated from the data X. This is often done by maximum likelihood estimation, and $\widehat{\theta}$ are the parameters that maximise the likelihood, given the chosen relationship between explanatory and outcome variables. Since only $\widehat{\theta}$ can be estimated from the data, we define the expected estimated K-L distance: $\mathbb{E}_Y\mathbb{E}_X\left[log\left(g(X|\widehat{\theta}(Y))\right)\right]$. [157] X and Y are random samples of the explanatory and outcome variables, respectively.

Akaike demonstrated that estimating the expected estimated K-L distance with the log-likelihood, $log\mathcal{L}$, is systematically biased, and that bias is approximately p, the number of parameters in g. There is the following relation between the relative expected estimated K-L distance and the maximised log-likelihood: [159]

$$log\mathcal{L}\left(\widehat{\theta}|X\right) - p = Constant - \widehat{\mathbb{E}}_f\left[log\left(g(X|\widehat{\theta})\right)\right] \tag{3.4}$$

From that relationship he defined the AIC, as follows: [159]

$$AIC = -2 * log\mathcal{L}\left(\widehat{\theta}|X\right) + 2 * p \tag{3.5}$$

As seen for the relative K-L distance, it is not the absolute value of AIC that matters but its relative value when compared to other AICs. The addition of known parameters will decrease the AIC, but any additional parameter that needs to be estimated from the data will incur a penalty on the AIC (p becomes $p+1$). When the addition of parameters in g induces very little knowledge on the unknown data generating mechanism f, there is over-fitting to the sample of data X and the AIC values begin to increase again. The increase in the likelihood does not compensate for the increase in p. The sample size N

of the data X limits the number of effects that can be estimated reliably and hence the capacity to reach the true generating distribution.

3.2.2 BIC

BIC stems from the Bayesian framework, although it remains valid beyond the Bayesian context. Along with other criteria, the BIC was developed with the a-priori idea that a true model exists. [160] It rests on several further assumptions: (1) the true model is contained in the pool of models tested, (2) the dimension (p) of that true model is relatively low (e.g. 1 to 5), and (3) that dimension is fixed and does not increase with increasing sample size. BIC was developed for prediction rather than to get a better understanding of the mechanisms under study. It is not an estimator of relative K-L distance. [157]

BIC is an approximation of the Bayes Factor, when the prior distribution of the parameters is the uniform distribution. The Bayes Factor is a ratio of likelihoods: the likelihood under the null hypothesis over the likelihood under the alternative hypothesis. The formula for the BIC is as follows:

$$BIC = -2 * log\mathcal{L}\left(\widehat{\theta}|X\right) + p * log(N) \tag{3.6}$$

The philosophies behind AIC and BIC diverge when we consider that the sample size N of the data X could increase. In the AIC philosophy, the true generating mechanism can be estimated with a larger number of parameters, when the sample size increases; in the BIC philosophy the true generating mechanism is set, whatever the information available. [160] When N is large, the target models are therefore different, when using AIC or BIC.

Burnham and Anderson [157] (section 6.4.5) show that AIC can be justified as a Bayesian model selection criterion, with a different set of prior probability distributions on the model set, rather than the uninformative uniform priors used by BIC. The prior distributions corresponding to the AIC are dependent on both N and p.

Bayesian model averaging aims to account for model selection uncertainty, by using the posterior probability of each model, given the data, as a weight in the estimation of the quantity of interest. A similar approach is proposed using AIC-weights for model averaging. [157]

3.3 Model selection using information criteria

As reported in Chapter 1, background knowledge is of paramount importance for selecting models, variables, and functional forms of effects. That knowledge will be used to code and make an initial selection of variables, to define the types of models, to deal with missing information when necessary, and to define what effects are allowed in the automatized model selection.

These initial steps and decisions remain in the context of prediction. First and foremost, background knowledge of both cohorts, (i) the cohort used for model building, and (ii) the cohort on which predictions are made, is crucial. For instance, there may be potential differences in the coding of variables, or in missingness mechanisms.

A second step is to make sure the data at hand provide enough information to estimate and summarise overall patterns of cancer survival, and that these can be confidently extrapolated to cohorts of patients that did not contribute to model building.

Next, one needs to decide on a modelling strategy. We present here an adaptation of a forward stepwise algorithm, originally proposed by Royston and Sauerbrei, [88, 89, 93] introduced and adapted to the relative survival data setting in Chapter 1. Rather than using likelihood ratio tests to select specific functional forms, interactions and presence of variables, we compare candidate models based on the values of their AICs or BICs. The model selection becomes a longer process, but the effects are selected in a hierarchical fashion.

3.4 Multi-model inference

Breiman [161] highlighted the 'quiet scandal' that ignores model-based uncertainty in inference when assuming that the selected model is the only one ever considered. When inference derives from regression models, there are many ways variables and effects are screened and selected, from background knowledge to specific algorithms. Such selection leads to models being discarded, and one final model to be selected as the model from which inference is derived. Inference is then derived as if we had come to the one selected model with certainty. Furthermore, it is not guaranteed that post-model selection inference tools such as p-values and 95% confidence intervals are valid: this has been recognised in the literature for many years (see list of the relevant literature in Berk et al. [162]. Several valid estimates of the standard errors, after any model selection techniques, may they be algorithmic, based on background knowledge of the field, informal, or ad-hoc, have been

proposed [162–165] and the research area is very active. Implementation of model averaging addresses that issue: [166] In multi-model inference, the focus turns to selecting (1) a set of models that will contribute to the estimations, and (2) the best way to estimate the weights each model parameter should be given when the models are pooled.

The AIC is used in Burnham and Anderson [157] to illustrate the use of multi-model inference. The idea behind multi-model inference is that the results of all models bearing equivalent support from the data are used in post-estimation of the quantities of interest. Averaging is the mean by which the model's parameters are combined.

The distance between the AIC values of two models are used to decide whether a model is close enough to the best performing model to be part of the multi-model inference. The best performing model is the model with the smallest of all AICs. As an analogy to the likelihood of parameters β given a set of data X and a model m, each model m is given a likelihood ℓ_m that it generated the data, given the data. This is calculated as the exponential of minus half the difference between the model's AIC, AIC_m, and the minimum AIC, AIC_0:

$$\ell_m(m|X) = exp\left(-\frac{1}{2}\triangle_m\right) \tag{3.7}$$

where $\triangle_m = AIC_m - AIC_0$.

The likelihood of the best model, that is the model with the minimum AIC, is 1, since $\triangle_0 = 0$ for that model. Several different models may have equivalently high likelihood to be K-L best model and there is no statistical reason to favour one over another. Therefore all M-best models are given weights, w_m, $m = 1 : M$. These are calculated as the ratio between the likelihood of a model ℓ_m over the sum of the likelihoods of all models contributing to the model-averaged estimates:

$$w_m = \frac{exp\left(-\frac{1}{2}\triangle_m\right)}{\sum_{k=1:M} exp\left(-\frac{1}{2}\triangle_k\right)} \tag{3.8}$$

Using the weights w_m and each of the M-best models' parameters $\widehat{\theta}_1 \ldots \widehat{\theta}_M$, we define the model-averaged parameters as $\overline{\overline{\theta}} = \sum_{k=1:M} w_k * \widehat{\theta}_k$. Model averaging means that the final, model-averaged, parameters used for inference correspond to the average of the parameters estimated for several models fitted on the same data and equally likely to have generated the data. [167]

Typically a model m is considered K-L best models when its AIC_m is within 2 of AIC_0, that is $\triangle_m \leq 2$. Multi-model inference using BIC values is identical, and restricts to most likely models, given the data.

In survival analysis and in complex generalised linear model settings, Burnham and Anderson [157] (p. 153) advise to combine the final outputs, such as the hazard at given times in the follow up, rather than the parameter estimates. They state:

'While it is often appropriate to average slope parameters in linear regression models, structural parameters in non-linear models should not be averaged'.

We consider that the baseline excess hazard and the effect of follow-up time are structural parameters. Therefore we proceed to model-average the predicted response, such as the excess hazard of death at given times, for given patient's characteristics. Uncertainty around model selection is taken into account in the final estimated functions via the calculation of the unconditional variance of the quantity of interest defined in Burnham and Anderson, [157] page 162.

$$\widehat{var}\left(\bar{\bar{\theta}}\right) = \left\{ \sum_{m=1}^{M} w_m * \sqrt{\widehat{var}\left(\hat{\theta}_m\right) + (\hat{\theta}_m - \bar{\bar{\theta}})^2} \right\}^2 \tag{3.9}$$

3.5 Multi-model inference for the prediction of cancer survival: manuscript in revision with Statistical Methods in Medical Research

In the following manuscript, in revision with Statistical Methods in Medical Research, we introduce the following concepts, novel to the field of modelling cancer survival:

- AIC- and BIC-model selection
- Multi-model inference
- Model-based prediction of cancer survival

The methods and rational for all three aspects are described in detail. Empirical data using historical cohorts of patients are used for a demonstration of their practical use.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	273792	**Title**	Mrs
First Name(s)	Camille		
Surname/Family Name	Maringe		
Thesis Title	On the prediction and projection of cancer survival		
Primary Supervisor	Prof. Bernard Rachet		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Statistical Methods in Medical Research
Please list the paper's authors in the intended authorship order:	Camille Maringe, Aurélien Belot, Bernard Rachet
Stage of publication	**Undergoing revision**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I am lead author on the paper. I designed the study in consultation with co-authors, I performed the analyses, interpreted the results and presented the results in the manuscript. I drafted the manuscripts and received comments from all co-authors.

SECTION E

Student Signature	[redacted]
Date	6/02/2020.

Supervisor Signature	[redacted]
Date	5 February 2020

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–18


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280220934501
journals.sagepub.com/home/smm
SAGE

Prediction of cancer survival for cohorts of patients most recently diagnosed using multi-model inference

Camille Maringe, Aurélien Belot and Bernard Rachet

Abstract

Despite a large choice of models, functional forms and types of effects, the selection of excess hazard models for prediction of population cancer survival is not widespread in the literature. We propose multi-model inference based on excess hazard model(s) selected using Akaike information criteria or Bayesian or Schwarz information criteria for prediction and projection of cancer survival. We evaluate the properties of this approach using empirical data of patients diagnosed with breast, colon or lung cancer in 1990–2011. We artificially censor the data on 31 December 2010 and predict five-year survival for the 2010 and 2011 cohorts. We compare these predictions to the observed five-year cohort estimates of cancer survival and contrast them to predictions from an a priori selected simple model, and from the period approach. We illustrate the approach by replicating it for cohorts of patients for which stage at diagnosis and other important prognosis factors are available. We find that model-averaged predictions and projections of survival have close to minimal differences with the Pohar-Perme estimation of survival in many instances, particularly in subgroups of the population. Advantages of information-criterion based model selection include (i) transparent model-building strategy, (ii) accounting for model selection uncertainty, (iii) no a priori assumption for effects, and (iv) projections for patients outside of the sample.

Keywords

Cancer survival, prediction, projection, multi-model inference, Akaike information criteria, Bayesian or Schwarz information criteria

1 Introduction

Cancer survival is a public health measure that complements the reporting of incidence, prevalence and mortality.[1] Projections of incidence and mortality figures are common practice.[2–5] These trends are often extrapolated to get estimates of the future burden of cancer for planning purposes, or based on scenarios reflecting the likely effect of new screening strategies, or changes in the distributions of risk factors.[6–8]

Survival models do not show good predictive performances.[9,10] This may be one of the reasons why prediction and projection of survival are, by far, less routinely made.

While prognosis research is focused on individual risk prediction scores,[11,12] we are interested here in predicting cancer survival for cohorts of patients as a whole or by reasonably large sub-groups, and we refer to these as population predictions. In that context, accurate individual-level predictions are less crucial since we intend to produce marginal estimates of survival. Many different survival models may be fitted to the data, and we focus here on regression models assuming multiplicative effects of explanatory variables on the hazard of death. A specificity of survival analysis is that the effects of variables may vary through follow-up time (time-dependent effect) and selecting the right effects can then be challenging. Background knowledge and model selection algorithms help narrow down the choice of models to the most appropriate one(s).[13]

Cancer Survival Group, London School of Hygiene & Tropical Medicine, Keppel Street, London, UK

Corresponding author:
Camille Maringe, Cancer Survival Group, London School of Hygiene & Tropical Medicine, Keppel Street, London, UK.
Email: Camille.Maringe@lshtm.ac.uk

When considered, model selection tends to be based on likelihood ratio tests,[14–16] using usually backward or forward selection strategy (or a combination of both). A single model is therefore selected as the best fit for the data or for subsequent prediction. We see two drawbacks to such approach. First, it means discarding effects that may have been equally likely to those selected. Second, once a model is chosen, no uncertainty relative to the selection is pertained to the model-based estimates and post-model selection inference.[17,18] In the context of prediction, we believe it is critical to consider that there may be several models equally likely to have generated the data. This is the philosophy of Bayesian model selection[19] and multi-model inference also described by Burnham and Anderson.[20] Lastly, hypothesis testing may perform poorly when using observational data,[20] as they are designed to detect discrepancies between the model and the data, rather than express how close a model is to the data: the larger the sample, the easier it is to detect (small) discrepancies.[21]

The Akaike Information Criteria (AIC)[22] is a likelihood-based measure that estimates the expected relative distance between the fitted model and the unknown true mechanism. AIC values can be compared between different, non-necessarily nested, models. Contrasting AIC values asymptotically coincides with generalised leave-one-out cross-validation.[23] The Bayesian or Schwarz Information Criteria (BIC) is an estimator of the Bayes Factor, aiming to quantify the evidence for one model against another.[24]

This article is organised as follows: the next section introduces the cancer registry data linked to electronic health records. The following section discusses the setting of relative survival for the estimation of cancer net survival,[25] the multi-model inference and the prediction tools used to assess the accuracy of the predicted estimates of net survival. Then, we present results on a historical, low-resolution, data setting for the prediction and projection of five-year survival for patients most recently diagnosed, to highlight the properties of the method. An application follows, based on more recent, high-resolution data including information on stage at diagnosis: a setting that motivates multivariable modelling and multi-model inference. The discussion highlights the advantages of multi-model inference and potential extensions conclude the manuscript.

2 Material

We use data of the population-based cancer registry of England. Virtually all cancer cases diagnosed in England are registered. Quality controls are performed at the time of registration, and prior to data analysis[26] to ensure there are no duplicate registrations and the sequence of dates (birth, diagnosis, latest vital status) is logical, among other checks.

We analyse records of adult patients (15–99 years) diagnosed with malignant lung cancer (men only, ICD-9: 162, ICD-10: C33-C34), breast cancer (women, ICD-9: 174, ICD-10: C50) or colon cancer (men only, ICD-9: 153, ICD-10: C18) in 1990 through to 2011. We define patients' information on socio-economic status based on their postcode of residence using the Townsend[27] and the income domain of the Index for Multiple Deprivation[28,29] scores for the years 1990–2000 and 2001–2011, respectively. Both scores are ecological and based upon responses to census questions relative to income and wealth, by small areas (Enumeration Districts until 2000 and Lower Super Output Areas from 2001). The areas are grouped by quintiles of area-level deprivation distribution, according to their score, from least (quintile 1) to most (quintile 5) deprived.

The latest vital status of patients is obtained from linking the cancer registrations to the mortality databases maintained by the Office for National Statistics. A vital status indicator is assigned to all patients together with a date of last known vital status, or death where appropriate. Patients are followed up until 31 December 2015.

Stage at diagnosis is one of the most important predictors of survival. It is based on the T (tumour size), N (lymph node involvement) and M (metastatic or not) components of the TNM stage at diagnosis classification.[30] Until recently, its recording, through combining information from pathology laboratories, hospital records, and Multidisciplinary Team records, was not complete or accurate for many cancers in population-based cancer registry data in England. High proportions of missing information on stage at diagnosis make it difficult to study its effect on survival through time.[31]

3 Methods

3.1 Scenarios studied

3.1.1 Low-resolution data setting: empirical evaluation of the properties of multi-model inference

We focus here on the cancers of colon (men), lung (men), and breast (women). First, we artificially restrict the follow-up to 31 December 2010. To compare the impact that varying numbers of cohorts have on the accuracy of the predictions, we run several model selections on cohorts of patients diagnosed in 1990–2010, or 1995–2010, or

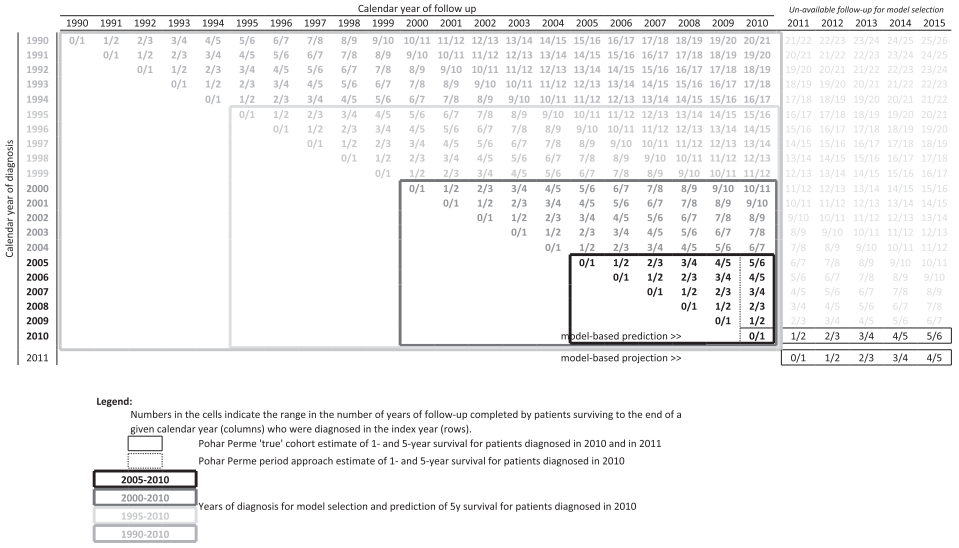


Figure 1. Structure of the data as used in the low-resolution data setting.

2000–2010, or 2005–2010. We predict excess hazard and five-year cancer survival for patients diagnosed in 2010, patients for whom only the first year of follow-up contributed to the model selection. We also project excess hazard and five-year cancer survival for patients diagnosed in 2011.

Since follow-up beyond 31 December 2010 is neither used in the estimation of the regression parameters nor in model selection, we are able to contrast the predicted five-year survival of these patients to their actual survival as observed until 31 December 2015 by group of patients and overall. Similarly, patients diagnosed in 2011 do not contribute to the modelling at all. Nonetheless we compare the results of their projections to their five-year survival as observed until 31 December 2015. Figure 1 summarizes how the data are used in this low-resolution data setting, highlighting what is supposed known and unknown, and the cohorts of patients used in model selection.

3.1.2 High-resolution data setting: illustration

We identify groups of patients for whom the proportion of missing stage at diagnosis is the lowest. For lung cancer, we select patients who were diagnosed at ages 50–74 between 2008 and 2012, and living in the East and North East of England (missing stage up to 14%).[32] For breast cancer, we analyse patients diagnosed at ages 50–84 in 2005–2011, living in the West Midlands (stage missing up to 12%).[33,34] For those two groups of patients, we can develop prediction models that include stage at diagnosis, as well as an indicator of mode of presentation (emergency for lung cancer, screening for breast cancer) and performance status (lung cancer). We predict lung and breast cancer survival up to four years after diagnosis for patients diagnosed in 2010 or 2011, for whom only the first year after diagnosis contributes to model selection and estimation of effects, and project cancer survival for patients diagnosed in 2011 and 2012.

3.2 Net survival

We aim to answer the following question: "What is the predicted *cancer* survival of cancer patients?" We focus on net survival, which measures survival among a defined cohort of cancer patients under the assumption that they only die of the studied cancer. This marginal survival measure is therefore independent of the deaths from other causes. Thus, this is the quantity of interest when aiming to compare cancer survival between countries and over time. Despite international classification, the determination of the cause of death is not standardised enough through time, or between registrars, for the cause of death to be used in our analyses. Hence, we aim to estimate cancer (net) survival in the relative survival setting using excess hazard models.[35,36] Several forms of models exist

exhibiting different ways of modelling the baseline excess hazard of death, and interactions with follow-up time.[37–45]

The main assumption of excess hazard models is that the observed mortality of the cohort of patients (λ) is the sum of two forces of mortality: the excess mortality hazard (λ_E), assumed to be the mortality hazard directly or indirectly due to cancer, and the expected or other causes mortality hazard, which is considered to be well approximated by the general population mortality hazard (λ_P).[46,47]

$$\lambda(t, \boldsymbol{x}) = \lambda_E\ (t, \boldsymbol{x}) + \lambda_P\ (a + t, y + t, \boldsymbol{z})$$

The cancer mortality hazard, λ_E, at time t for given patient's covariates $\boldsymbol{x}$, such as age at diagnosis (a) and calendar year of diagnosis (y), is what we need to estimate. We derive mortality due to causes other than cancer at time t by population tables of mortality, defined for the population from which cancer patients come from, i.e. with similar features (age at time t: $a + t$, calendar year at time t: $y + t$, sex, levels of deprivation, geographical area of residence, ethnicity when possible, etc., summarised in $\boldsymbol{z}$, a subset of patient's covariates $\boldsymbol{x}$).

First, we use the non-parametric Pohar-Perme (PP) estimator,[48] a consistent estimator of net survival, to obtain cancer survival for patients diagnosed in 2010 with follow-up until 2015. That estimator relies on the observed and expected proportions of patients still alive at each time of event. Patients may die of other causes, thus preventing their cancer survival time to be observed. The cohort of patients therefore changes structurally throughout follow-up time and is not representative of the original cohort of patients. An inverse-probability-of-censoring weighting is applied to adjust for this informative censoring, so that the contribution of each patient to the estimator is weighted by the inverse of the probability that the patient is expected to survive until each time of event (using population tables of mortality). The period approach PP estimator is also used to predict survival for patients diagnosed in 2010, using information from patients diagnosed in previous cohorts, alive in 2010, with potential follow-up until 31 December 2010 ('period approach').[49] The period approach derives survival in a similar fashion to life expectation.

Second, we use flexible, multivariable models, to estimate excess mortality hazard $\lambda_E(t, \boldsymbol{x})$, individual ($S_{E,i}(t, \boldsymbol{x})$) and cohort ($S_E(t)$) net survival.[50] The logarithm of the baseline excess hazard is modelled using restricted cubic spline functions, with three degrees of freedom, that is two internal knots (located at the tertiles of the event time distribution) and two boundary knots

$$\log(\lambda_0(t)) = \gamma_0 + \gamma_1 B_1(t) + \gamma_2 B_2(t) + \gamma_3 B_3(t)$$

where the spline basis functions $B_i(t),\ i = 1,2,3$, are derived from the knots.[51]

Time-dependent effect of each variable is included using an interaction between each variable and the logarithm of time since diagnosis. As an example, the equation of the model is as follows, given two prognostic variables x_1, continuous, and x_2, categorical (with J categories, $j = 1, \ldots J$)

$$\lambda_E\ (t, \boldsymbol{x}) = \lambda_0(t) * \exp\left(\beta_1(t) * f(x_1) + \sum_{j=2}^{J} \beta_{2,j}(t) * I_{x_2=j}\right)$$

where $f(x_1) = x_1$ if the effect of x_1 on the logarithm of excess mortality is linear, and $f(x_1)$ is a spline function when the effect of x_1 is not linear, while $\beta_1(t) = \beta_1$ if the effect of x_1 is proportional, and $\beta_1(t) = \beta_1 * \log(t)$ if not; the same applies to $\beta_{2,j}(t)$.

We use the Stata commands *stns*[52] to implement the PP cohort and period approaches, and *strcs*[53] for fitting the flexible parametric models.

3.3 Model selection

We present two specific model-selection algorithms here, but wish to highlight that any other sound algorithm could be used. We adapt *mfpigen*, the model-selection algorithm designed by Royston and Sauerbrei, including tests for interactions,[54] and our adaptation of *mvrs*[14] for interactions[13] using the Akaike information criteria (AIC)[22] and the Bayesian Information Criteria (BIC).[55] AIC is one of the criteria designed to express the 'distance'

between two models,[20] that is an estimate of the distance between our model and the model that did generate the data. AIC is defined from the log-likelihood of the model, $\mathcal{L}$ and its number of parameters, p.

The log-likelihood of the excess hazard models fitted here is

$$\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{x},\ \delta_i,\ t_i) = \sum_{i=1}^{N} \delta_i \log\{\lambda_P(t_i,\boldsymbol{z}) + \lambda_E(t_i,\boldsymbol{\beta},\boldsymbol{x})\} + \log\{S_E(t_i)\}$$

such that

$$AIC = -2*\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{x},\ \delta_i,\ t_i) + 2*p$$

AIC can be shown to be equivalent to a likelihood ratio test multiplied by a constant, meaning that there is an associated positive probability (p-value) that it rejects the null hypothesis, when it is true. That p-value is 0.157 when models are nested and differing by $1df$.[56,57]

BIC comes from a consistent class of criteria. It does not estimate the distance between the true model and the model under consideration, but aims to consistently point to the true model even when sample size increases, if the true model is part of the models considered. Its value varies with the number of parameters p and the number of events d.

$$BIC = -2*\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{x},\ \delta_i,\ t_i) + p*\log(d)$$

The Royston and Sauerbrei algorithm is a succession of likelihood ratio tests comparing two models at a time in a logical sequential order. The algorithm starts by fitting the simplest model to the data, using linear and proportional effects of all variables. Starting with the most significant effect (i.e. lowest p-value), more complex versions of the effects of each variable are tested, one at a time, such as non-linearity and time-dependency.

Our adaptation follows the same logical steps, but the models' AICs or BICs are compared, two at a time. If the lowest criterion is over two digits away from the larger criterion, the model pertaining to the larger criterion is discarded. If both models have criteria within two of each other, both models are kept, and more complex models derived from each of these are further compared. A rational for the choice of a difference of 2 is provided in section 3.4 using evidence ratios.

The original Royston and Sauerbrei algorithm yields one single model, from which all inference about measure of effects, associations, and outcome prediction is derived. Our proposed algorithm based on Information criteria leads to the selection of several models, which are equally likely to have generated the data, given their AIC or BIC are within 2 digits of the minimum AIC.

3.4 Multi-model inference (model averaging)

In the following, XIC is used to stand for AIC or BIC, interchangeably. From the multiple models selected (i.e. models having similar support from the data), say M, we need to combine the M estimates to obtain one estimate of the excess hazard from which we derive the cohort cancer survival. Since the models are selected using XIC, each has a known XIC from which we derive XIC-weights as follows:

Let us define the model with lowest XIC (XIC_{min}) as m_{min}. We define the distance between m_{min} and any other model m, $\Delta_m = XIC_m - XIC_{min}$, and the likelihood of model m given the data is $\mathcal{M}(m|x) = \exp\left(-\frac{1}{2}*\Delta_m\right)$.[20]

The weights w_m of each of the M models m reflect how much evidence there is for model m being the actual model that generated the data. Weights are defined such that they sum to 1, $\sum_{m=1}^{M} w_m = 1$

$$w_m = \frac{\mathcal{M}(m|x)}{\sum_{m=1}^{M} \mathcal{M}(m|x)}$$

'Evidence ratios' for a model m versus model n are defined as the ratio of their weights w_m and w_n, as

$$e(m,n) = \frac{w_m}{w_n} = \frac{\exp\left(-\frac{1}{2}*\Delta_m\right)}{\exp\left(-\frac{1}{2}*\Delta_n\right)}$$

If we suppose model m is the model with minimum XIC, we have

$$e(m,n) = \frac{1}{\exp\left(-\frac{1}{2}*\Delta_n\right)} = \exp\left(\frac{1}{2}*\Delta_n\right)$$

Therefore, we see exponential increase in evidence for the model with minimal XIC with increased distance to that XIC. The evidence ratio between models m and n is 2.7 if $\Delta_n = 2$ (and 7.4 and 54.6 when $\Delta_n = 4$ or 8, respectively). This is a rational for selecting models with XIC within two digits of the minimum XIC, where the evidence for m versus n is not so strong.

Given the potential complexity of the effects on the excess mortality hazard, we average the quantity modelled rather than the parameter estimates.[20] This is specifically advised in Burnham and Anderson: "Structural parameters in non-linear models should not be averaged" and model averaging should rather be done on "the predicted expected response variable $\hat{E}(y)$".[20] Therefore, the XIC-weights are used to combine the model-based individual excess hazards estimated at each time t after diagnosis. Borrowing from the reasoning of both the algorithmic model-selection[16,58] and the multi-model inference literature, we follow the steps below to combine the model-based estimates into a model-averaged estimate:

a. Run XIC-based algorithm (e.g. *mfpigen* or the adapted *mvrs*)
b. Isolate the M best models
c. Calculate the XIC-weights, w_m, for each model m, $(m = 1, \ldots, M)$
d. From the estimated parameters, derive the excess mortality hazards, at pre-defined times t after diagnosis (e.g. monthly) for each model m: $\hat{\lambda}_{i,m}(t)$ for each individual i with covariates x_i in the data.
e. Calculate the model-average excess mortality hazard (for patient i) at each pre-defined time t, such that

$$\hat{\lambda}_{i,MA}(t) = \sum_{m=1}^{M} w_m * \hat{\lambda}_{i,m}(t)$$

The model-average cumulative excess hazard, $\hat{\Lambda}_{i,MA}$ may easily be obtained as well

$$\hat{\Lambda}_{i,MA}(t) = \sum_{m=1}^{M} w_m * \hat{\Lambda}_{i,m}(t)$$

f. If the quantity of interest is cohort net survival, we first calculate individual model-averaged net survival, $\hat{S}_{i,MA}(t) = \exp(-\hat{\Lambda}_{i,MA}(t))$ at each time t. Then, we estimate cohort net survival by averaging the individual net survival values, $\hat{S}_{MA}(t) = \frac{1}{N}\sum_{i=1}^{N} \hat{S}_{i,MA}(t)$ at time t.

The unconditional variance estimator of the model-averaged estimate is derived in Burnham and Anderson (pp.158–164)[20] and follows earlier work presented in Buckland et al.[17] We adapted this derivation to our setting where we averaged the predicted expected response variable (i.e. the excess mortality hazards). The variance estimator for the model-averaged outcome combines the XIC-weights, w_m, and the estimated variances of each individual model estimates, $\widehat{var}(\hat{\lambda}_{i,m}(t))$, such that

$$\widehat{var}\left(\hat{\lambda}_{i,MA}(t)\right) = \left\{\sum_{m=1}^{M} w_m * \sqrt{\widehat{var}(\hat{\lambda}_{i,m}(t)) + \left(\hat{\lambda}_{i,m}(t) - \hat{\lambda}_{i,MA}(t)\right)^2}\right\}^2$$

This estimator has components of within-variation ($\widehat{var}(\hat{\lambda}_{i,m}(t))$) and between variation $\left(\hat{\lambda}_{i,m}(t) - \hat{\lambda}_{i,MA}(t)\right)^2$, thus quantifying the uncertainty with regards to model selection. This unconditional variance estimator assumes perfect pairwise correlation between $\hat{\lambda}_{i,m}(t) - \hat{\lambda}_{i,MA}(t)$ and $\hat{\lambda}_{i,n}(t) - \hat{\lambda}_{i,MA}(t)$, as derived from models m and n. This leads to a conservative variance estimate, i.e. the estimated variance tends to be too large.[17]

3.5 Checking predictions

After selecting a (or a set of) best model(s), predicting our outcome of interest, and averaging the outcomes, we are interested in quantifying the distance between these estimates and the observed cancer survival of the patients. Since we have follow-up information until 31 December 2015, we estimate net survival of patients diagnosed in 2010 and 2011 using the PP non-parametric estimator of cancer survival (see point 3.2. above).

We aim to compare our predictions to what will be estimated in the future, given the data available then. We recognise that the PP estimator of survival, often used for policy making and planning, is a consistent estimator of net survival, but cannot be regarded as the 'truth'.

To quantify the difference between the population-based prediction using our model-average estimate and the PP net survival estimates, we define the Root Mean Integrated Square Difference (RMISD) of prediction. This measure contrasts the predicted survival to the estimated PP survival and we approximate this quantity using G groups defined by age group and deprivation quintile (low-resolution data) or by age group, stage and deprivation (high-resolution data)

$$RMISD = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\int (\hat{S}_g(u) - S_g(u))^2 \mathrm{d}u}$$

S_g is the non-parametric PP estimate of cancer survival for group g, while $\hat{S}_g$ is the prediction of survival for the same group of patients using (i) model-averaging, or (ii) a simple model or (iii) the period approach. The integral is approximated using the Gauss-Legendre quadrature with 20 nodes. We choose to calculate RMISD for survival measured at one and five years after diagnosis.

4 Results

4.1 Low-resolution data setting: empirical evaluation of the properties of multi-model inference

4.1.1 Description of the data

Between 1990 and 2010, there were an average of 18,233 and 8,636 men diagnosed with lung and colon cancer, respectively, and 32,493 women diagnosed with breast cancer, every year. The number of cancer patients was multiplied by at least 1.5 for breast (women) and colon (men), but slightly decreased for lung cancer (men). Five-year net survival increased for all cancers between 1990 and 2010, with the largest increase for lung cancer, from 5.3% in 1990 to 9.0% in 2010 (online Appendix Table 1).

4.1.2 Model selection

The functional forms of the selected variables are displayed in Table 1 (columns 1-4), along with the XIC of the selected model(s) (column 5), the XIC-distance to the model with the closest XIC (column 7), and where appropriate, XIC-weights (column 6). Adding earlier cohorts to patients diagnosed in 2005–2010 hardly change the functional form selected for the effects of age, year of diagnosis or deprivation, as well as their interactions, especially when using the *mfpigen* algorithm with AIC, or when using BIC (with either algorithm). More complex models (including time-dependent effects of the interaction between age and deprivation) are selected by our adapted algorithm using AIC: these include time-dependent age-deprivation interactions (breast cancer) and age-deprivation and year-deprivation interactions (lung cancer). With BIC selection, there is almost no difference in the complexity of the models selected by *mfpigen* and our adapted algorithm: the models selected are identical for colon and lung cancers, and the only selected interactions differ for breast cancer. To contrast with the models selected, we also apply a simple model with all variables modelled with a linear (when continuous), proportional hazard effect on excess mortality, in each of the four cohorts of patients. The XIC values of these simple models (Table 1) are constantly higher than that of the selected models except for the AIC values of the lung cancer models selected by our adapted algorithm.

4.1.3 Root mean integrated square difference for the prediction of net survival

Root mean integrated square difference (RMISD) is measured throughout the first five years after diagnosis. By group defined by age and deprivation level, we calculated Integrated Square Differences (ISD) (see formula of the

Table 1. Models selected following model selection algorithms, by cancer and cohort of patients used in model selection.

		Age (A)	Year of diagnosis (Y)	Deprivation (D)	Interactions	XIC[a]	XIC[a] weights (%)	XIC[a] distance with next selected model	RMISD at 5 years 2010[b]	RMISD at 5 years 2011[c]
Breast cancer										
Algorithm adapted for interactions										
AIC	2005–2010	NL TD	NL TD	TD	A*D (TD), Y*D	205,554.5		18	0.030	0.028
	2000–2010	NL TD	NL TD	TD	A*D (TD), Y*D	589,478.3		61	0.032	0.032
	1995–2010	NL TD	NL TD	TD	A*D (TD), Y*D (TD)	1,072,626.0	69		0.032	0.033
	1995–2010	NL TD	NL TD	TD	A*D (TD), Y*D	1,072,628.0	31	104		
	1990–2010	NL TD	NL TD	TD	A*D (TD), Y*D (TD)	1,605,374.0		12	0.039	0.042
BIC	2005–2010	NL TD	NL TD	PH		205,742.0		18.3	0.026	0.026
	2000–2010	NL TD	NL TD	TD	Y*D	589,840.6		10.4	0.033	0.032
	1995–2010	NL TD	NL TD	TD	Y*D	1,073,076.0		406.0	0.033	0.033
	1990–2010	NL TD	NL TD	TD	Y*D (TD)	1,605,849.0		53.4	0.039	0.042
mfpigen algorithm										
AIC	2005–2010	NL TD	NL TD	TD	A*D	205,576.9			0.024	0.027
	2000–2010	NL TD	NL TD	TD	A*D	589,540.4			0.031	0.034
	1995–2010	NL TD	NL TD	TD	A*D	1,072,766.0			0.030	0.035
	1990–2010	NL TD	NL TD	TD	A*D	1,605,607.0			0.039	0.045
BIC	2005–2010	NL TD	NL TD	PH		205,742.0		18.3	0.026	0.026
	2000–2010	NL TD	NL TD	TD		589,766.8		8.6	0.031	0.033
	1995–2010	NL TD	NL TD	TD		1,073,035.0		13.8	0.030	0.034
	1990–2010	NL TD	NL TD	TD	A*D	1,605,901.0		9.5	0.039	0.045
Simple models[d]										
	2005–2010	L PH	L PH	PH		208,763.6			0.069	0.075
	2000–2010	L PH	L PH	PH		598,469.7			0.079	0.087
	1995–2010	L PH	L PH	PH		1,087,460.8			0.088	0.098
	1990–2010	L PH	L PH	PH		1,625,430.8			0.089	0.098
Period approach										
Colon cancer										
Algorithm adapted for interactions										
AIC	2005–2010	NL TD	L TD	TD		97,225.4	65		0.094	0.117
	2005–2010	NL TD	NLTD	TD		97,226.7	35	4		
	2000–2010	NL TD	NL TD	TD	A*D (TD), Y*D	215,214.2		39	0.097	0.110
	1995–2010	NL TD	NL TD	TD	Y*D	338,571.2		192	0.093	0.108
	1990–2010	NL TD	NL TD	TD	A*D (TD), Y*D	455,678.2		85	0.095	0.110
BIC	2005–2010	NL TD	NL TD	PH		97,379.7		637.0	0.095	0.116
	2000–2010	NL TD	NL TD	TD		215,438.8		1,483.9	0.093	0.108
	1995–2010	NL TD	NL TD	TD		338,747.7		171.3	0.091	0.109

(continued)

Table 1. Continued.

									RMISD at 5 years	
		Age (A)	Year of diagnosis (Y)	Deprivation (D)	Interactions	XIC[a]	XIC[a] weights (%)	XIC[a] distance with next selected model	2010[b]	2011[c]
	1990–2010	NL TD	NL TD	TD		455,968.5		3,315.6	0.093	0.111
mfpigen algorithm										
AIC	2005–2010	NL TD	NL TD	TD	A*D	97,225.2	67		0.090	0.115
	2005–2010	NL TD	NL TD	TD		97,226.7	33			
	2000–2010	NL TD	NL TD	TD	A*D	215,253.6			0.094	0.111
	1995–2010	NL TD	NL TD	TD	A*D	338,556.8			0.090	0.109
	1990–2010	NL TD	NL TD	TD	A*D	455,763.4			0.092	0.112
BIC	2005–2010	NL TD	NL TD	PH		97,379.7		67.3	0.095	0.116
	2000–2010	NL TD	NL TD	TD		215,438.8		41.2	0.093	0.108
	1995–2010	NL TD	NL TD	TD		338,747.7		72.7	0.091	0.109
	1990–2010	NL TD	NL TD	TD		455,968.5		66.9	0.093	0.111
Simple models[d]										
	2005–2010	L PH	L PH	PH		98,608.6			0.175	0.170
	2000–2010	L PH	L PH	PH		217,925.7			0.155	0.150
	1995–2010	L PH	L PH	PH		342,308.5			0.147	0.138
	1990–2010	L PH	L PH	PH		460,688.7			0.141	0.138
Period approach										
Lung cancer										
Algorithm adapted for interactions										
AIC	2005–2010	NL TD	NL TD	TD	Y*D	110,543.9		410	0.122	0.142
	2000–2010	NL TD	NL TD	TD	A*D, Y*D	220,096.1		31	0.103	0.128
	1995–2010	NL TD	NL TD	TD	A*D, Y*D	323,089.4		1,041	0.102	0.127
	1990–2010	NL TD	NL TD	TD	A*D, Y*D	418,484.9		1,386	0.102	0.126
BIC	2005–2010	NL TD	NL TD	PH		101,805.6		573.1	0.120	0.148
	2000–2010	NL TD	NL	PH		202,668.9		932.1	0.107	0.125
	1995–2010	NL TD	NL TD	PH		290,457.4		423.2	0.105	0.129
	1990–2010	NL TD	NL TD	TD		378,957.9		193.5	0.105	0.127
mfpigen algorithm										
AIC	2005–2010	NL TD	NL TD	TD	A*D	101,644.7			0.112	0.147
	2000–2010	NL TD	NL TD	TD	A*D	202,427.4			0.102	0.128
	1995–2010	NL TD	NL TD	TD	A*D	290,194.5			0.101	0.125
	1990–2010	NL TD	NL TD	TD	A*D	378,710.8			0.100	0.126
BIC	2005–2010	NL TD	NL TD	PH		101,805.6		69.8	0.120	0.148
	2000–2010	NL TD	NL	PH		202,668.9		75.6	0.107	0.125
	1995–2010	NL TD	NL TD	PH		290,457.4		63.1	0.105	0.129

(continued)

Table 1. Continued.

									RMISD at 5 years	
		Age (A)	Year of diagnosis (Y)	Deprivation (D)	Interactions	XIC[a]	XIC[a] weights (%)	XIC[a] distance with next selected model	2010[b]	2011[c]
	1990–2010	NL TD	NL TD	TD		378,957.9		55.0	0.105	0.127
Simple models[d]										
	2005–2010	L PH	L PH	PH		102,492.2			0.119	0.113
	2000–2010	L PH	L PH	PH		203,852.5			0.119	0.113
	1995–2010	L PH	L PH	PH		292,678.4			0.125	0.117
	1990–2010	L PH	L PH	PH		381,722.1			0.124	0.117
Period approach									0.076	

L: Linear; NL: non-linear; TD: time-dependent; PH: proportional hazard; *: interaction; AIC: Akaike information criteria; BIC: Bayesian or Schwarz information criteria.
[a]XIC stands for AIC or BIC depending on model.
[c]RMISD are calculated by averaging the integrated square differences measured in each of the age and deprivation groups.
[d]Prediction of five-year survival for patients diagnosed in 2010 and 2011.
[d]A model with all variables modelled with a linear proportional effect.

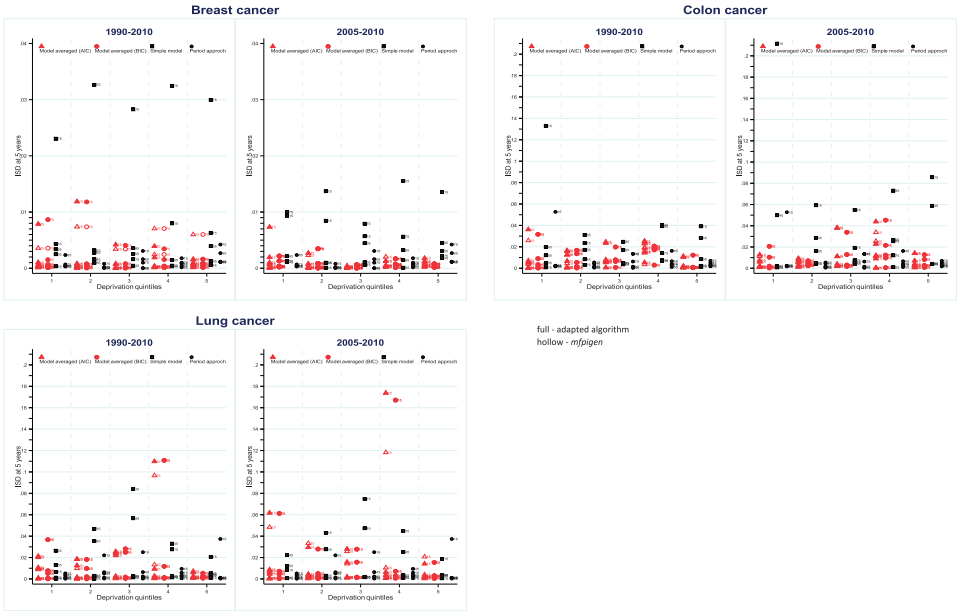


Figure 2. Integrated Square Difference (ISD) between NS predicted by each AIC or BIC model-averaged, simple models, and the period approach, compared to the PP cohort survival for patients diagnosed in 2010, by age group, deprivation from 1990 to 2010 and from 2005 to 2010 cohorts of patients used in model selection.

RMISD in section 3.5.) between model-averaged net survival estimates and the PP estimates using known follow-up until 31 December 2015 for patients diagnosed in 2010 (Figure 2 for the cohorts 1990–2010 and 2005–2010, and online Appendix Figure 1 for all four cohorts), and for patients diagnosed in 2011 (online Appendix Figure 2).

Breast cancer: All ISDs are very small, and largest differences are seen for the oldest age-group when survival is predicted by the simple model. Similar observations can be made for the projection of survival for patients diagnosed in 2011, not included in model selection (online Appendix Figure 2). The more recent the cohorts of patients, the better the estimates of survival: ISDs are smaller when using 2005–2010 cohorts only versus 1990–2010 cohorts.

Colon cancer: Simple models lead to high ISD for different age and deprivation groups, such as patients aged 15–54 years in the most deprived group and 45–54 years in deprivation quintile 4. Except for patients aged 15–44 in the least deprived group, 2010-period approach estimates show low ISD. ISD values remain stable and low, whatever the number of cohorts used in multivariable model-averaged prediction of survival (online Appendix Figure 1). ISDs for patients aged 15–44 and 45–54 years are slightly higher when the models are used for projection of survival for patients diagnosed in 2011.

Lung cancer: Except for patients aged 15–44, in deprivation quintile 4, for whom model-averaged ISD is large, model-averaged ISD are generally lower than ISD derived from the simple model, and smaller or similar to most of the 2010-period ISDs. Model-averaged predictions for patients aged 15–44, in deprivation quintile 3 and diagnosed in 2011 show very high ISD. Such large ISDs are also observed, but to a lesser extent, for simple model estimates (online Appendix Figure 2).

The highlighted patterns in ISD, for all three cancers, are observed for (i) AIC (triangular shapes) and BIC (circular shapes) selected models, and following (ii) model selection using *mfpigen* (hollow red symbols) and our adapted algorithm (full red symbols).

By averaging the ISD values displayed in Figure 1 and online Appendix Figures 1 and 2, the RMISD values summarise the overall differences in the survival curves (Table 1). For all cancers, model-averaged estimates of survival lead to the smallest RMISD, in comparison to using pre-defined simple models (Table 1). Nonetheless, there are differences between cancers: for breast cancer, there is a small advantage in restricting the model selection and estimation to the cohorts of patients diagnosed in the last five years, while for patients diagnosed with colon or lung cancer, longer time-trends yield better estimates of survival for patients for whom follow-up is

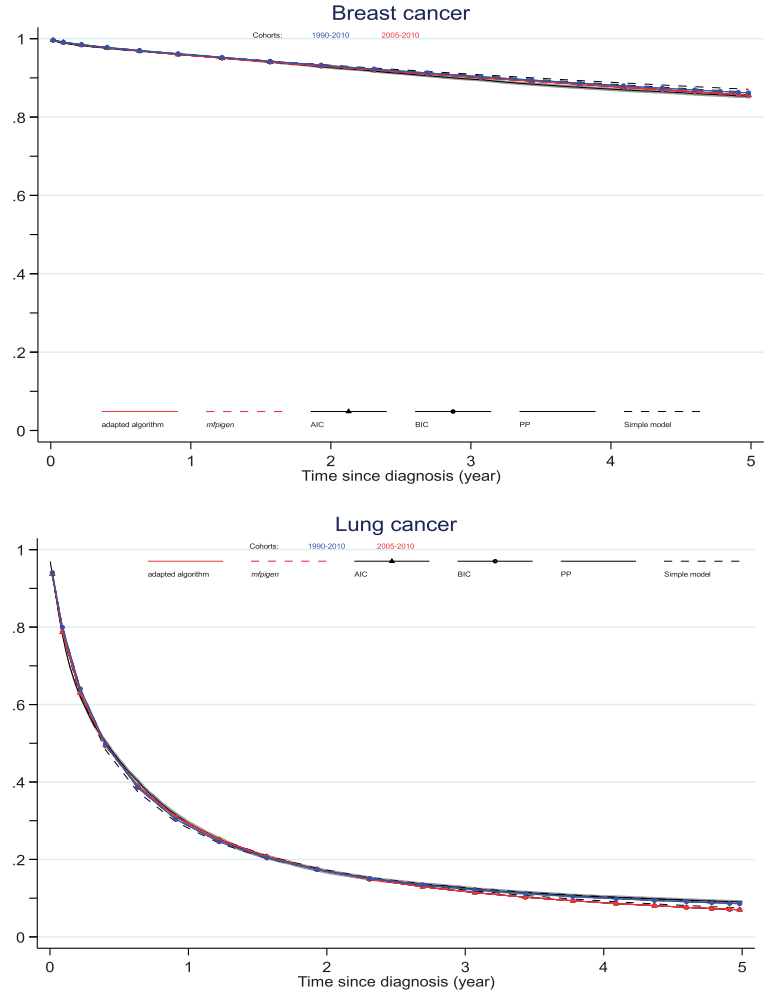


Figure 3. Net survival curves: comparison between the PP estimates and estimates from 1990 to 2010 and from 2005 to 2010 cohorts of patients used in model-averaging from AIC and BIC model selection.

not yet available. The simple models yield the highest RMISD values, for each cancer and each cohort, except for lung cancer when using the AIC-based multi-model inference from the adapted *mvrs*. (Table 1)

Figure 3 shows what the actual differences are on the overall cohort net survival curves, contrasting cancer survival estimated from the simple model, and from model-averaged selection, to the 2010-cohort approach. The differences between the model-averaged estimates of survival up to five years are tiny when contrasting AIC and BIC selection, adapted *mvrs* or *mfpigen* algorithm. Nonetheless, they do reflect the conclusions from RMISD: additional cohorts of patients are necessary for a better prediction of lung cancer survival. Net survival estimated from model selection and when necessary, model averaging, are closer to the PP estimates than estimates from simple models.

4.2 High-resolution data setting: illustration

This illustration rests on richer datasets to allow inclusion of the effects of potentially key prognostic factors such as stage at diagnosis, mode of presentation (emergency presentation for lung cancer, screening for breast cancer) and performance status (lung cancer) on the excess hazard.

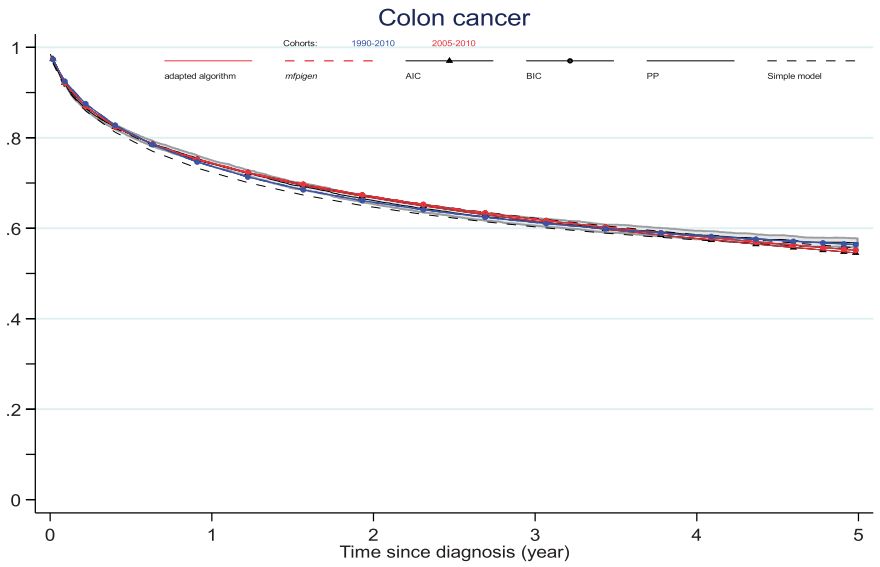


Figure 3. Continued

For both cancers, between 1 and 10 models have similar support from the data, given their AIC, but only one model given its BIC, when restricting to models with BICs within two of each other: we report the effects estimated, the AIC or BIC and corresponding weights in Table 2. The models selected to model breast cancer survival have AIC weights between 16.1% and 42.3% (adapted *mvrs*) and between 6.6% and 17.8% (*mfpigen*); the models selected to model lung cancer survival have AIC weights between 11.6% and 29.4% (adapted *mvrs*). The selected models with the highest AICs are only just over two units away from the next model: 2.2 for breast (both algorithms) and lung (*mfpigen*), but 316.3 units away for lung (adapted R&S). The effects of deprivation (PH) and stage (TD) for breast cancer, and the effects of stage (TD), performance status (TD), emergency presentation (TD), and an interaction between age and deprivation for lung cancer, are selected in all models.

Model-averaged estimates of the excess hazard are presented in online Appendix Figure 3, highlighting differences between these and those estimated by simple models, especially for stage IV with larger excess hazard estimated with the simple models.

There is very little difference between the AIC (*mfpigen* and adapted *mvrs*) and BIC model-averaged survival curves for patients diagnosed with breast or lung cancer (Figure 4). Survival estimated from the simple model, although modelling the effects of all variables, does differ for both cancers, especially at stages IV (breast) and II and III (lung).

The confidence intervals around the net survival curves highlight uncertainty related to data sparsity but also model selection.

5 Discussion

We contrast the predictions and projections of cancer survival derived from a model-averaged approach and (i) a-priory simple model and (ii) non-parametric period approach. We use an algorithm for model selection that has been used in cancer epidemiology,[13] for scanning methodically through the possible effects of independent factors on the excess hazard of death, merely to illustrate the multi-model inference in cancer survival. Indeed, any other algorithm based on screening through possible effects could have been adapted to the information criteria paradigm. We implement the model averaging methodology for the selection of the best model(s)[20] using AIC and BIC as selection criteria. We show the lethality and the rate of improvement of cancer survival determine how many past cohorts of patients are needed to predict and project survival with best accuracy. We also show that allowing for multi-model estimation of cancer survival generally results in restricted mean integrated square differences as good as or better than the non-parametric period approach. In some cases, despite larger AICs

Table 2. Models selected, and associated AIC, BIC for the prediction and projection of four-year cancer survival.

	Age (A)	Year of diagnosis (Y)	Deprivation (D)	Stage (S)	Performance status (PS, lung)	Presentation: screening (breast) emergency (lung)	Interactions	XIC[b]	XIC[b] weights (%)	XIC[b] distance with next selected model
Breast cancer										
(adapted)	L TD	NL TD	PH	TD		PH		9,481.6	23.4	
	L TD	NL	PH	TD		PH		9,480.4	42.3	
	L TD	NL TD	PH	TD		TD		9,482.1	18.2	
	L	NL	PH	TD		TD		9,482.4	16.1	2.2
(mfpigen)	L TD	NL	PH	TD		PH		9,480.4	17.8	
	NL TD	L	PH	TD		PH	A*D	9,481.1	13.1	
	NL TD	L	PH	TD		PH		9,480.7	15.7	
	L TD	NL TD	PH	TD		PH		9,481.6	9.9	
	NL	L	PH	TD		TD	A*D	9,482.1	7.8	
	NL	L	PH	TD		TD		9,482.1	7.7	
	L TD	NL TD	PH	TD		TD		9,482.1	7.7	
	L	NL	PH	TD		TD		9,482.4	6.8	
	NL	L	PH	TD		PH	A*D	9,482.3	7.0	
	NL	L	PH	TD		PH		9,482.4	6.6	2.2
BIC-selection (adapted & mfpigen)	L	NL	PH	TD		PH		9,576.3[c]		
Simple model[a]	L	L	PH	PH		PH		9,619.2		
Lung cancer										
(adapted)	L	NL	TD	TD	TD	TD	A*D	7,820.2	15.5	
	L	NL TD	PH	TD	TD	TD	A*D	7,819.0	29.4	
	L	NL	PH	TD	TD	TD	A*D	7,819.1	26.8	
	L	NL TD	TD	TD	TD	TD	A*D	7,820.1	16.6	
	L TD	NL TD	PH	TD	TD	TD	A*D	7,820.8	11.6	316.3
(mfpigen)	NL	L TD	PH	TD	TD	TD	A*D	7,819.9		2.2
BIC-selection (adapted & mfpigen)	L	NL	PH	TD	TD	TD		7,967.4[c]		
Simple model[a]	L	L	PH	PH	PH	PH		8,369.8		

L: Linear; NL: non-linear; TD: time-dependent; PH: proportional hazard; *: interaction; AIC: Akaike information criteria; BIC: Bayesian or Schwarz information criteria.

[a]A model with all variables modelled with a linear proportional effect.

[b]XIC stands for AIC or BIC depending on model.

[c]BIC value.

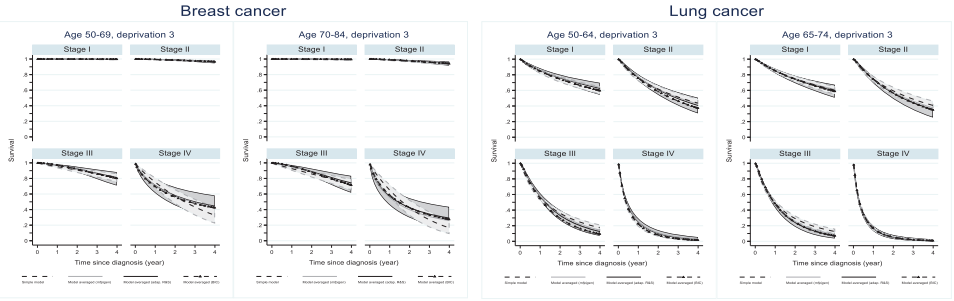


Figure 4. Up to four-year net survival for patients diagnosed with breast or lung cancer by age and stage for patients in the third deprivation quintile.

or BICs, simple models produced accurate predictions, similar to model-averaged predictions, but projections from these models do not estimate cancer survival as well.

There are many advantages to estimating survival using IC-based model selection and multi-model inference. (1) Transparent model building strategy: the algorithm walks through the effects of variables in a hierarchical and systematic fashion. (2) Uncertainty relative to model selection is taken into account in the variance of the estimated outcomes. (3) There is no assumption that an effect is simple, without checking it can or needs to be simple. (4) Projections for patients outside of the training sample are possible, which is not possible using period approach.

The results show that for breast cancer patients, only patients diagnosed in the five years prior to the year for which we need to make five-year survival predictions are needed to produce accurate predictions and projections. This can be explained by survival increasing at constant pace of about 3–8% per five years in the last 20 years. By contrast, for lung and colon cancers, cancer survival increased irregularly in the last 20 years: close to 30% increase in five-year lung cancer survival between 2005 and 2010, but no increase between 1990 and 1995 and similarly, 12% increase in colon cancer survival between 2005 and 2010 but only 3% between 2000 and 2005. More cohorts of patients are needed to predict and project five-year survival accurately, due to these irregular trends in survival. These considerations need to be borne in mind when using the most recent cohorts for the prediction of cancer survival.

Bayesian, cross-validation and bootstrap-based approaches are also likely to perform well in excess hazard model selection. Nonetheless these carry high computational demands. BIC readily links with Bayesian model averaging and is asymptotically consistent in estimating the true generating model.[59] Despite AIC asymptotically equivalent to cross-validation,[23] and therefore a tool of choice for model selection in the context of prediction, it tends to overestimate the dimension of the true model.[59] Furthermore, multi-model inference has theoretical and practical advantages, particularly for predictions.[20] **[AQ1]** These include: (1) taking into account uncertainty in model selection, leading to more robust results as they do not necessarily depend upon a particular model; (2) choosing to average models that have AIC within two of the minimum AIC should help keep the number of considered models reasonable; (3) model averaging avoids one to have to defend the choice of model, it makes convincing stakeholders from different backgrounds and highlighting the robustness of the results easier.[60] We recognise the limitation that model uncertainty remains conditional on the model set, as all models come from a unique model set.[61] Other approaches which have proved useful for predictions would broaden the model sets considered (e.g. LASSO, Random Survival Forest) and could provide interesting research developments but would need to be adapted to the relative survival data setting.

We focus here on predicting and projecting five-year net survival, as most events happen in the short term following a cancer diagnosis, certainly for colon and lung cancers. By contrast, breast cancer patients experience long-term excess mortality. Therefore, we performed additional analyses for the prediction of 10-year breast cancer survival. We found that a model-averaged 10-year survival prediction leads to a smaller difference than from a simple model or from a non-parametric period approach (data not shown).

Using empirical data in the low-resolution setting, we can only compare our predictions and projections to the consistent non-parametric PP estimates of cancer survival, for patients diagnosed in a given year. We acknowledge that these remain estimates of survival, rather than the "truth", but we argue that they will be what is produced when the follow-up information becomes available, to contrast trends in cancer outcomes.[62–64] Nonetheless, both

non-parametric and parametric outcomes are estimating the same quantity since the models are adjusting for the variables that constitute the strata of the PP estimates.

However, in the application, due to data sparsity by strata defined by the values of the variables adjusted for in the models, it was not possible to compare the model-averaged estimates of net survival to the PP. Indeed, when the PP is not stratified by the same prognostic factors, it is not estimating the same quantity as the model-based estimates. The results of the high-resolution setting are presented to motivate the use of multi-model inference for the prediction and projection of cancer survival. The differences between predictions derived from a simple model versus IC-based approach, however, highlight that it would be relevant to conduct such comparison in a larger population in which variables such as stage at diagnosis, mode of presentation and performance status are available.

Multi-model inference, as presented here, allows model parameters to remain the raw information for the estimation of each model's outcome of interest. Such outcomes are then averaged, and interpretation of the predictions can only be made on the outcome. Multi-model inference increases the ability to perform better predictions while retaining interpretability of the averaged outcomes.[61,65] It seems to be a good compromise between best-model selection strategy (high interpretability but poor predictions) and ensemble learning strategy (high predictions but poor interpretability). For patients and their carers, prediction of the remaining survival time represents their main interest. However, this point estimate of time carries poor predictive capability.[9] Hence, much of the literature focuses on prediction of survival probabilities, at individual or population level. In the field of prognosis research at individual level, there is a growing emphasis on improving the quality of published risk scores so they are useful to individual patient prognosis.[12]

Here, we aim to predict and project population-based levels of survival, rather than individual cancer survival predictions. It is the reason why we do not rely on standard loss functions, or usual measures of discrimination and calibration. It is still important to gather accurate information on the main prognostic factors, and make sure models are correctly specified since correct model specification and availability of individual patient characteristics improve prediction. All of this is exemplified in both scenarios here, low- and high-resolution data settings, in which complex prognosis models are compared.

6 Conclusion

We recommend that, given a set of variables that may influence levels of cancer mortality, possible excess hazard models should be assessed systematically. We encourage analysts to consider that a model may not be singled out as the best model. Model averaging using Kullback-Leibler distance such as AIC, or Bayesian principles such as BIC, allows users to consider several equivalent models and effects, and to take account of the uncertainty relative to model selection in the estimation of the variance of the outcomes. Prediction and projection of cancer survival can best be done using such carefully selected parametric models.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Camille Maringe https://orcid.org/0000-0002-8739-9565

Supplemental material

Supplemental material for this article is available online.

References

1. Ellis L, Woods LM, Estève J, et al. Cancer incidence, survival and mortality: explaining the concepts. 2014; **135**: 1774–1782. DOI: doi:10.1002/ijc.28990. [AQ2]
2. Smittenaar CR, Petersen KA, Stewart K, et al. Cancer incidence and mortality projections in the UK until 2035. *Br J Cancer* 2016; **115**: 1147–1155.

3. Maddams J, Utley M and Moller H. Projections of cancer prevalence in the United Kingdom, 2010-2040. *Br J Cancer* 2012; **107**: 1195–1202.
4. Olsen AH, Parkin DM and Sasieni P. Cancer mortality in the United Kingdom: projections to the year 2025. *Br J Cancer* 2008; **99**: 1549–1554. 2008/10/16. DOI: 10.1038/sj.bjc.6604710.
5. Mistry M, Parkin DM, Ahmad AS, et al. Cancer incidence in the United Kingdom: projections to the year 2030. *Br J Cancer* 2011; **105**: 1795–1803. 2011/10/29. DOI: 10.1038/bjc.2011.430.
6. Friedenreich CM, Barberio AM, Pader J, et al. Estimates of the current and future burden of cancer attributable to lack of physical activity in Canada. *Prevent Med* 2019; **122**: 65–72.
7. Poirier AE, Ruan Y, Grevers X, et al. Estimates of the current and future burden of cancer attributable to active and passive tobacco smoking in Canada. *Prevent Med* 2019; **122**: 9–19.
8. Liu Z, Jiang Y, Fang Q, et al. Future of cancer incidence in Shanghai, China: Predicting the burden upon the ageing population. *Cancer Epidemiol* 2019; **60**: 8–15.
9. Henderson R, Jones M and Stare J. Accuracy of point predictions in survival analysis. *Stat Med* 2001; **20**: 3083–3096.
10. Graf E, Schmoor C, Sauerbrei W, et al. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999; **18**: 2529–2545.
11. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med* 2015; **13**: 1.
12. Riley RD, van der Windt D, Croft P, et al. *Prognosis research in healthcare: concepts, methods, and impact*. Oxford, UK: Oxford University Press, 2019.
13. Maringe C, Belot A, Rubio FJ, et al. Comparison of model-building strategies for excess hazard regression models in the context of cancer epidemiology. *BMC Med Res Methodol* 2019 (in press). [AQ3]
14. Royston P and Sauerbrei W. Multivariable modeling with cubic regression splines: a principled approach. *Stata J* 2007; **7**: 45–70.
15. Royston P and Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Stat Med* 2004; **23**: 2509–2525.
16. Wynant W and Abrahamowicz M. Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Stat Med* 2014; **33**: 3318–3337.
17. Buckland ST, Burnham KP and Augustin NH. Model selection: an integral part of inference. 1997. [AQ4]
18. Efron B. Estimation and accuracy after model selection. *J Am Stat Assoc* 2014; **109**: 991–1007.
19. Rossell D and Rubio FJ. Additive Bayesian variable selection under censoring and misspecification. 2019. [AQ5]
20. Burnham KP and Anderson DR. *Model selection and multimodel inference: a practical information-theoretic approach*. New York, NY: Springer Science & Business Media, 2003.
21. Raftery AE. Choosing models for cross-classifications. *Am Sociol Rev* 1986; **51**: 145–146.
22. Akaike H. A new look at the statistical model identification. *Selected Papers of Hirotugu Akaike*. Springer, 1974, pp.215–222. [AQ6]
23. Stone M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. 1977; **39**: 44–47. [AQ7]DOI: 10.1111/j.2517-6161.1977.tb01603.x.
24. Kass RE and Raftery AE. Bayes factors. *J Am Stat Assoc* 1995; **90**: 773–795.
25. Pohar Perme M, Esteve J and Rachet B. Analysing population-based cancer survival - settling the controversies. *BMC Cancer* 2016; **16**: 933.
26. Li R, Abela L, Moore J, et al. Control of data quality for population-based cancer survival analysis. *Cancer Epidemiol* 2014; **38**: 314–320.
27. Jarman B, Townsend P and Carstairs V. Deprivation indices. *Br Med J* 1991; **303**: 523–523.
28. Department for Communities and Local Government. *The English indices of deprivation 2007*. 2008. London. [AQ8]
29. Department for Communities and Local Government. *The English indices of deprivation 2010*. 2011. London. [AQ9]
30. Sobin LH, Gospodarowicz M and Wittekind C. *TNM classification of malignant tumours*. 7th ed. New York: John Wiley & Sons, 2009.
31. Benitez-Majano S, Fowler H, Maringe C, et al. Deriving stage at diagnosis from multiple population-based sources: colorectal and lung cancer in England. *Br J Cancer* 2016; **115**: 391–400.
32. Belot A, Fowler H, Njagi EN, et al. Association between age, deprivation and specific comorbid conditions and the receipt of major surgery in patients with non-small cell lung cancer in England: a population-based study. *Thorax* 2019; **74**: 51–59.
33. Morris M, Woods LM and Rachet B. What might explain deprivation-specific differences in the excess hazard of breast cancer death amongst screen-detected women? Analysis of patients diagnosed in the West Midlands region of England from 1989 to 2011. *Oncotarget* 2016; **7**: 49939–49947.
34. Woods LM, Morris M and Rachet B. No 'cure' within 12 years of diagnosis among breast cancer patients who are diagnosed via mammographic screening: women diagnosed in the West Midlands region of England 1989-2011. *Ann Oncol* 2016; **27**: 2025–2031.
35. Esteve J, Benhamou E, Croasdale M, et al. Relative survival and the estimation of net survival: elements for further discussion. *Stat Med* 1990; **9**: 529–538.

36. Mariotto AB, Noone AM, Howlader N, et al. Cancer survival: an overview of measures, uses, and interpretation. *J Natl Cancer Inst Monogr* 2014; **2014**: 145–186.
37. Rubio FJ, Remontet L, Jewell NP, et al. On a general structure for hazard-based regression models: an application to population-based cancer research. *Stat Meth Med Res* 2018: 962280218782293. [AQ10]
38. Charvat H, Remontet L, Bossard N, et al. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Stat Med* 2016; **35**: 3066–3084.
39. Remontet L, Bossard N, Belot A, et al. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Stat Med* 2007; **26**: 2214–2228.
40. Remontet L, Uhry Z, Bossard N, et al. Flexible and structured survival model for a simultaneous estimation of non-linear and non-proportional effects and complex interactions between continuous variables: performance of this multidimensional penalized spline approach in net survival trend analysis. *Stat Meth Med Res* 2018: 962280218779408. [AQ11]
41. Lambert PC and Royston P. Further development of flexible parametric models for survival analysis. *Stata J* 2009; **9**: 265–290.
42. Fauvernier M, Roche L, Uhry Z, et al. Multi-dimensional penalized hazard model with continuous covariates: applications for studying trends and social inequalities in cancer survival. *J Royal Stat Soc Ser C* 2019; **68**: 1233–1257.
43. Pohar Perme M, Henderson R and Stare J. An approach to estimation in relative survival regression. *Biostatistics* 2009; **10**: 136–146.
44. Cortese G and Scheike TH. Dynamic regression hazards models for relative survival. *Stat Med* 2008; **27**: 3563–3584.
45. Mahboubi A, Abrahamowicz M, Giorgi R, et al. Flexible modeling of the effects of continuous prognostic factors in relative survival. *Stat Med* 2011; **30**: 1351–1365.
46. Cutler SJ and Ederer F. Maximum utilisation of the life table method in analyzing survival. *J Chronic Dis* 1958; **8**: 699–712.
47. Ederer F, Axtell LM and Cutler SJ. The relative survival: a statistical methodology. *Natl Cancer Inst Monograph* 1961; **6**: 101–121.
48. Pohar Perme M, Stare J and Esteve J. On estimation in relative survival. *Biometrics* 2012; **68**: 113–120.
49. Brenner H and Gefeller O. An alternative approach to monitoring cancer patient survival. *Cancer* 1996; **78**: 2004–2010.
50. Belot A, Ndiaye A, Luque-Fernandez MA, et al. Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data. *Clin Epidemiol* 2019; **11**: 53–65.
51. Rutherford MJ, Crowther MJ and Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *J Stat Computat Simulat* 2015; **85**: 777–793.
52. Clerc-Urmès I, Grzebyk M and Hédelin G. Net survival estimation with stns. *Stata J* 2014; **14**: 87–102.
53. Bower H, Crowther MJ and Lambert PC. strcs: A command for fitting flexible parametric survival models on the log-hazard scale. *Stata J* 2016; **16**: 989–1012.
54. Royston P and Sauerbrei W. Multivariable modeling with cubic regression splines: a principled approach. *Stata J* 2007; **7**: 25.
55. Schwarz G. Estimating the dimension of a model. *Ann Statist* 1978; **6**: 461–464.
56. Steyerberg EW, Eijkemans MJC, Harrell FE Jr, et al. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. 2000; **19**: 1059–1079. [AQ12]DOI: 10.1002/(sici)1097-0258(20000430)19:8<1059::Aid-sim412>3.0.Co;2-0.
57. Heinze G, Wallisch C and Dunkler D. Variable selection – a review and recommendations for the practicing statistician. *Biometric J* 2018; **60**: 431–449.
58. Sauerbrei W, Royston P and Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat Med* 2007; **26**: 5512–5528.
59. Bozdogan H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* 1987; **52**: 345–370.
60. Hoeting JA, Madigan D, Raftery AE, et al. Bayesian model averaging: a tutorial. *Stat Sci* 1999; **14**: 382–401.
61. Lavoue J and Droz PO. Multimodel inference and multimodel averaging in empirical modeling of occupational exposure levels. *Ann Occup Hyg* 2009; **53**: 173–180.
62. Magadi W, Exarchakou A, Rachet B, et al. *Cancer survival in England: patients diagnosed between 2010 and 2014 and followed up to 2015*. 2016. [AQ13]
63. Exarchakou A, Rachet B, Belot A, et al. Impact of national cancer policies on cancer survival trends and socioeconomic inequalities in England, 1996-2013: population based study. *BMJ* 2018; **360**: k764.
64. Belot A, Coleman MP, Magadi W, et al. *Geographic patterns of cancer survival in England Adults diagnosed 2003 to 2010 and followed up to 2015*. 2017. Newport, Wales: Office for National Statistics.
65. Burnham KP and Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. *Sociol Meth Res* 2004; **33**: 261–304.

3.6 Model averaging: on what scale should we average?

As highlighted in section 3.4, the modelling of excess hazard of death leads to averaging model outcomes rather than model parameters. In the manuscript of section 3.5, we concentrate on averaging the individual excess hazard of death. This outcome measure is the natural outcome of excess hazard models. Excess mortality is a key measure of the burden of cancer on mortality at given times after follow up, directly extracted from the excess hazard models. It helps interpret trends and patterns of its survival counterpart: the estimates of net survival which are often used for tracking progress of cancer control through time. Both these measures are calculated in the hypothetical setting in which there is no competing burden of other causes of death, and therefore make the assumption that patients are immune to other causes of death. Individual excess mortality values represent an elementary component from which one can also derive crude probability of cancer death. [2]

There are many possible ways of reporting and communicating predictions from excess hazard models, and the choice for one or the other mostly depends on what the intended audience may be. [2] There is no theoretical reasons for model-averaging the excess hazard of death, as a substitution for other quantities. Indeed other quantities of interest can be averaged directly: this section highlights on-going work aiming to describe the steps leading to averaging other such quantities directly from the excess hazard models.

Let us define the following notations, needed in the subsequent paragraphs:

$\lambda_{Ei}^{m}(t)$ is the excess hazard of death for patient i estimated at time t by model m.

$\overline{\overline{\lambda}}_{Ei}(t)$ is the model-averaged excess hazard of death for patient i estimated at time t.

$\overline{S}_{E}(t)$ is the cohort net survival estimated at time t, from the individual $S_{Ei}(t)$.

$S_{Ei}^{m}(t)$ is the net survival for patient i estimated at time t by model m.

$\overline{\overline{S}}_{Ei}(t)$ is the model-averaged net survival for patient i estimated at time t.

$\overline{\overline{S}}_{E}(t)$ is the cohort net survival estimated at time t, from the model-averaged $\overline{\overline{S}}_{Ei}(t)$.

$CPD(t)$ is the crude probability of death at time t, also known as the cumulative incidence function in the classical competing risks setting. An individual crude probability of death measure, $CPD_i(t)$, can be calculated, as well as one coming from a specific model m, $CPD_i^m(t)$. The cohort model-averaged crude probability of cancer death is denoted $\overline{\overline{CPD}}(t)$.

The data introduced in the high-resolution setting of the manuscript in section 3.5 is used in the following sub-sections to illustrate the model averaging principles on other scales. These data refer to patients diagnosed with lung cancer, at ages 50-74 years between 2008 and 2012, and living in the East and North East of England and patients diagnosed with breast cancer, at ages 50-84 years in 2005-2011, living in the West Midlands. The variables available for prediction include age at diagnosis, deprivation, stage at diagnosis, as well as an indicator of mode of presentation (emergency for lung cancer, screening for breast cancer) and performance status (lung cancer).

3.6.1 Averaging on the excess hazard scale

In the manuscript (section 3.5), we predict individual excess hazard for each patient i at each time of interest t, by each selected model m: $\lambda_{Ei}^m(t)$. Given the weights w_m for each model m, $m = 1 : M$, we apply the formula for model averaging these estimates, such that:

$$\overline{\overline{\lambda}}_{Ei}(t) = \sum_{m=1:M} w_m * \lambda_{Ei}^m(t) \tag{3.10}$$

Given the relationship between excess hazard and net survival at individual level, the individual net survival estimates for each patient i at each time t are obtained such that:

$$\overline{\overline{S}}_{Ei}(t) = exp\left(-\int \overline{\overline{\lambda}}_{Ei}(u)du\right)$$

and the cohort net survival are

$$\overline{\overline{S}}_{E}(t) = \frac{1}{N}\sum_{i=1:N} \overline{\overline{S}}_{Ei}(t)$$

By re-arranging the expression of $S_{Ei}(t)$ we get an expression of these values from the original model-based values of individual net survival $S_{Ei}^m(t)$:

$$\begin{aligned}
\overline{\overline{S}}_{Ei}(t) =& exp\left(-\int_0^t \sum_{m=1}^{M} w_m * \lambda_{Ei}^m(u)du\right) \\
=& \prod_{m=1:M} exp\left(-\int_0^t w_m * \lambda_{Ei}^m(u)du\right) \\
=& \prod_{m=1:M} exp\left(-\int_0^t \lambda_{Ei}^m(u)du\right)^{w_m} \\
=& \prod_{m=1:M} S_{Ei}^m(t)^{w_m}
\end{aligned}$$

$$\overline{\overline{S}}_E(t) = \frac{1}{N} \sum_{i=1:N} \left(\prod_{m=1:M} S_{Ei}^m(t)^{w_m} \right) \tag{3.11}$$

In the context of averaging on the excess hazard, model-averaged cohort net survival are the average (over individuals) of the product of each model-based individual net survival to the power of their respective weight w_m.

3.6.2 Averaging on the net survival scale

Net survival is a probability, it is therefore conceptually easier to apprehend model averaging since values are bounded between 0 and 1. It is equivalent to take the average of individual net survival values, or the average of the cohort net survival directly. It corresponds to taking the average of a weighted sum, or the weighted sum of an average.

We predict individual net survival for each patient i at each time of interest t, by each selected model m : $S_{Ei}^m(t)$. The formula for model averaging these estimates from M models, each with an IC-weight w_m is simply:

$$\overline{\overline{S}}_{Ei}(t) = \sum_{m=1:M} w_m * S_{Ei}^m(t) \tag{3.12}$$

We can then derive cohort (or sub-group) net survival estimates such that:

$$\overline{\overline{S}}_E(t) = \frac{1}{N} \sum_{i=1:N} \overline{\overline{S}}_{Ei}(t) = \frac{1}{N} \sum_{i=1:N} \sum_{m=1:M} w_m * S_{Ei}^m(t) \tag{3.13}$$

From the unconditional variance estimator proposed by Burnham and Anderson [157] (page 162), the formula for the variance of $\overline{\overline{S}}_{Ei}(t)$ is given by

$$\widehat{var}\left(\overline{\overline{S}}_{Ei}(t)\right) = \left\{ \sum_{m=1}^{M} w_m * \sqrt{\widehat{var}\left(S_{Ei}^m(t)\right) + \left(S_{Ei}^m(t) - \overline{\overline{S}}_{Ei}(t)\right)^2} \right\}^2 \tag{3.14}$$

$\widehat{var}\left(S_{Ei}^m(t)\right)$ is obtained using the Delta method and the variance of the cohort net survival estimate:

$$\widehat{var}\left(\overline{\overline{S}}_E(t)\right) = \widehat{var}\left(\frac{1}{N} \sum_{i=1:N} \overline{\overline{S}}_{Ei}(t)\right) = \frac{1}{N^2} \sum_{i=1:N} \widehat{var}\left(\overline{\overline{S}}_{Ei}(t)\right) \tag{3.15}$$

We provide in table 3.1 the Restricted Mean Integrated Square Differences (RMISD) when model-averaging is performed on the excess hazard scale (formula (1)), compared to model-averaging of the net survival values directly (formula (2)). Using the high-resolution data,

Table 1 shows there is virtually no difference between the final estimated net survival curves. The differences between the estimated curves and the Pohar Perme (PP) are identical to the third digit place, whatever the size of the groups (split by age group, deprivation quintile and stage at diagnosis or deprivation and stage, or only deprivation quintiles). The corresponding net survival curves by age group, deprivation quintiles and stage are identical.

RMISD are calculated as the average ISD measured in each group defined by age, deprivation and stage, or deprivation and stage, or deprivation alone. Reducing the number of groups on which integrated square differences (ISD) are compared increases the sample size for each group, and therefore the stability of the PP cohort estimates (the standard to which all other estimation is compared).

3.6.3 Averaging on the crude probability of death scale

Crude probabilities of cancer death, also known as cumulative incidence function, report the probabilities of dying of cancer, in the presence of competing mortality due to causes of death other than cancer. [2] This is a measure defined in the 'real world' in which patients may experience other causes of death. Crude probabilities of death from cancer are most commonly calculated using the formula 3.16, at individual level CPD_i, and then for the cohort CPD as the average of all individual CPD_i:

$$CPD_i(t) = \int_0^t S_{Oi}(u) * \lambda_{Ei}(u)du \tag{3.16}$$

$$CPD(t) = \frac{1}{N}\sum_{i=1}^{N} CPD_i(t) = \frac{1}{N}\sum_{i=1}^{N}\int_0^t S_{Oi}(u) * \lambda_{Ei}(u)du \tag{3.17}$$

Where S_{Oi} is the overall survival and λ_{Ei} is the excess hazard for patient i. It makes sense that, for individual i, their probability of dying of cancer at time t is function of surviving all causes of death until t, and the hazard of dying from cancer at t. The crude probability is a cumulative function of time.

Given we do not observe much or any follow-up for the patients for whom we would like to predict a crude probability of death value, we cannot estimate their overall survival probabilities or excess hazard of death from their own data. λ_{Ei} is estimated following model selection and model-averaging as seen in formula 3.10. Similar to modelling excess hazard, we perform model selection and model-averaging of overall survival to estimate S_{Oi}. We detail the required steps:

Table 3.1: RMISD for Breast and Lung cancers: comparison of model-averaging on individual excess hazard *vs.* individual net survival quantities

	RMISD at 4 years [a]					
Model-averaged	Individual excess hazard (formula (1))			Individual net survival (formula (2))		
	age, depr., stage	depr., stage	depr.	age, depr., stage	depr., stage	depr.
Breast cancer						
(adapted)	0.13811	0.11300	0.02016	0.13815	0.11304	0.02015
(mfpigen)	0.13618	0.11150	0.02176	0.13648	0.11184	0.02170
Simple model [b]	0.13945	0.11531	0.02508	0.13945	0.11531	0.02508
Lung cancer						
(adapted)	0.18558	0.10916	0.04041	0.18547	0.10902	0.03992
(mfpigen)	0.18523	0.11114	0.03734	0.18523	0.11114	0.03734
Simple model	0.18903	0.11243	0.03773	0.18903	0.11243	0.03773

[a] RMISD are calculated as the average ISD measured in each group defined by age, deprivation and stage, or deprivation and stage, or deprivation alone. Reducing the number of groups on which integrated square differences (ISD) are compared increases the sample size for each group, and therefore the stability of the PP cohort estimates (the standard to which all other estimation is compared).

[b] A model with all variables modelled with a linear proportional effect

(i) Model selection for the estimation of overall survival using information criteria

(ii) Predict overall survival, by patient characteristics, from each of the models selected

(iii) Derive model-averaged individual survival $\overline{S}_{Oi}(t)$, using weights based on the information criteria chosen

(iv) Proceed to the calculation of the model-averaged CPD_i, using the model-averaged overall survival:

Using multi-model inference, we propose to use the model-averaged individual excess hazard $\overline{\lambda}_{Ei}(t)$ as an estimator of λ_{Ei} and $\overline{S}_{Oi}(t)$ an estimator of $S_{Oi}(t)$.

$$\overline{\overline{CPD}}_i(t) = \int_0^t \overline{S}_{Oi}(u) * \overline{\lambda}_{Ei}(u)du = \int_0^t \overline{S}_{Oi}(u) * \sum_{m=1:M} w_m * \lambda_{Ei}^m(u)du \tag{3.18}$$

And

$$\overline{\overline{CPD}}(t) = \frac{1}{N}\sum_{i=1}^{N} CPD_i(t) = \frac{1}{N}\sum_{i=1}^{N}\int_0^t \overline{S}_{Oi}(u) * \sum_{m=1:M} w_m * \lambda_{Ei}^m(u)du \tag{3.19}$$

In summary,

$$\overline{\overline{CPD}}(t) = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M} w_m * \int_0^t \overline{S}_{Oi}(u) * \lambda_{Ei}^m(u)du \tag{3.20}$$

Throughout the first four years of follow-up, we contrast the model-averaged crude probability of death curves, based on the two model selection algorithms presented in the manuscript in section 3.5, to the simple model's (blue line).

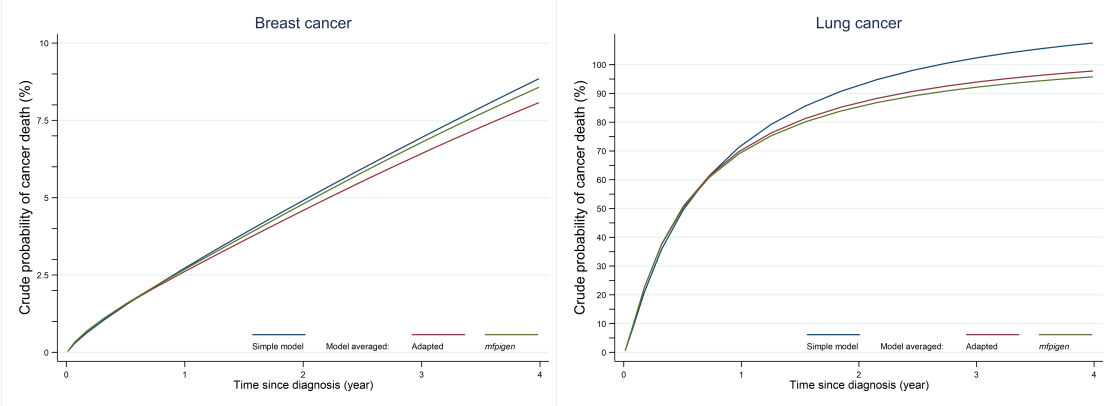


Figure 3.1: Crude probability of death as estimated from a simple model, or from multi-model inference.

The values obtained with the model-averaged estimates of CPD are as expected for these cancers, given the restriction to the 50-85 years and 50-75 years at diagnosis for breast and lung cancer, respectively (Figure 3.1). These graphs should be completed with the actual estimated crude probability of death, to check the accuracy of the model-based predictions.

3.7 Averaging measures of explained variation

In this section, we provide a taste for how we could apprehend validation of multi-model inference. Chapter 2 provides a review of measures that are used for evaluating the quality of model-based predictions. There, we also describe how we adapt a measure of explained variation to the context of modelling the excess hazard of death, using weights, defined as the probability that, given a patient's death, it is due to cancer. Here, we would like to provide a model-averaged version of that measure of explained variation.

From each selected model we can calculate the variation in outcome that is explained by the model. The measure of explained variation is based on ranks, defined by individually predicted values of excess hazard measured at each time of event for each patient i still at risk. In this context, one needs to average the individual excess hazard values. Since we average the models' outcomes into one final multi-model outcome, it is of interest to study what is the variation explained by that specific combination of models. We followed the steps below to average the excess hazard estimates, and measure the explained variation of the model-averaged excess hazards:

1. Identify the M models that are AIC-equivalent, and all equally susceptible to have generated the data.

For each time of event t^* until end of follow up:

1. Predict the individual excess hazard λ_{Ei}^m from each model m.
2. Average the values of λ_{Ei}^m to obtain a model-average estimate of the individual excess hazards, $\overline{\overline{\lambda}}_{Ei}$.
3. Rank the individual $\overline{\overline{\lambda}}_{Ei}$ and compare the ranks to the rank from null and perfect models, as described in Chapter 2.

The measure of explained variation, REw, is calculated cumulatively from diagnosis, until the end of follow-up; this is $REw(t)$. A time varying measure, the local REw, can also be derived using the information of 20 event times around each index time.

We compare below the values of explained variation for each selected model m, the model-averaged (both algorithms as described in 3.5) and the simple model. The simple model only contains linear and proportional effects of each of the prognostic factors on the excess hazard of death (see manuscript in section 3.5). The explained variation is calculated for the cohort of patients that only contributed up-to a year of follow-up to model-selection. We are therefore in the context of explained variation for the prediction of survival.

The explained variation of the breast cancer models are incredibly high (Figure 3.2). This is almost certainly due to the large selection of variables available, including stage at diagnosis and mode of presentation, which are very strong prognostic factors and will determine treatment strategies and ultimately survival. We notice the increase in explained variation following model-averaging (darker shades). Both the $RE(t)$ and local REw decline throughout the four years of follow-up.

As seen in the application of the manuscript presented in Chapter 2 (section 2.6.1), explained variation for the cohort of lung cancer patients is around 50% with a steep decline in the first eighteen months and then a stable level of explained variation, due to the cumulative nature of $RE(t)$ (Figure 3.3). The local RE, measured using a window of 20 events around each event, is steeply decreasing too with very little variation explained by the model-averaged effects beyond 24 months. Simple and two individual models selected using the adapted R&S show local REw below 0 after 18 months. Higher explained variation is seen for model-averaging.

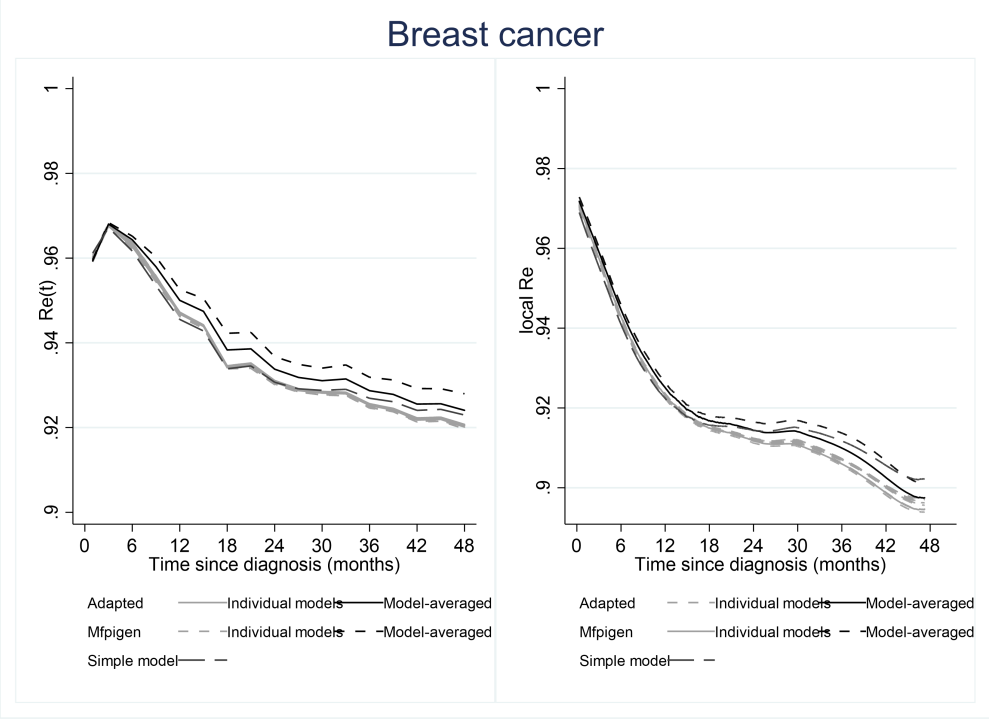


Figure 3.2: Time-varying and local REw, patients diagnosed with breast cancer in 2010

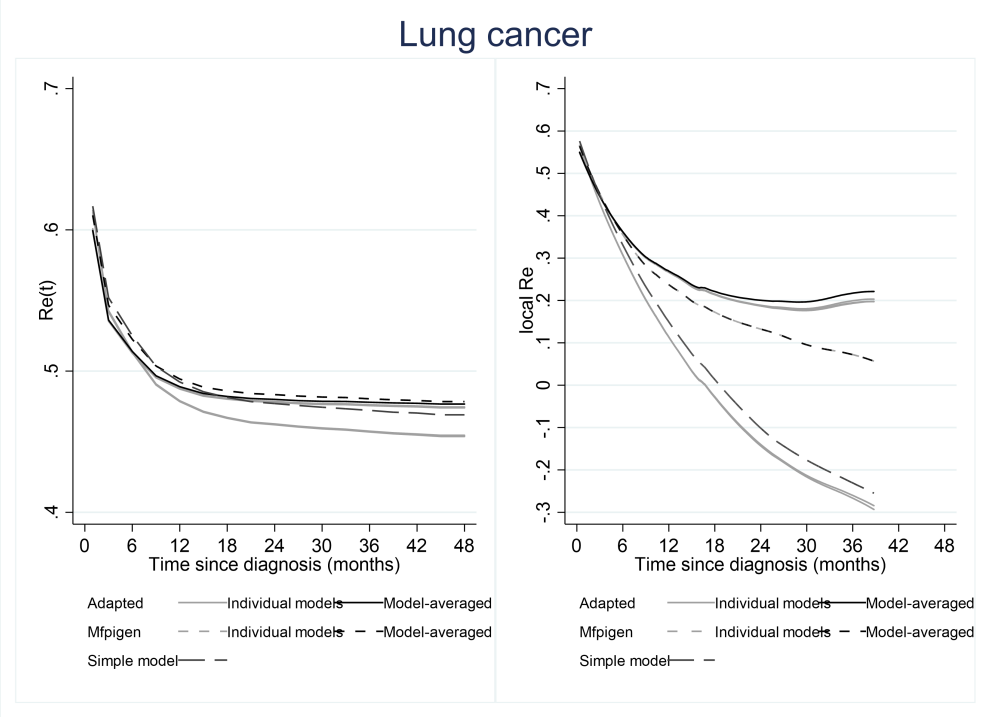


Figure 3.3: Time-varying and local REw, patients diagnosed with lung cancer in 2011

3.8 Discussion

In this Chapter, we explore the concept of multi-model inference for better prediction and projection of cancer survival. We aim to show the benefits of information criteria-based model selection, and of model averaging. We show good performance of multi-model inference, when compared to using a unique simple model for the prediction and projection of cancer survival. This is true in both a low- and high-resolution setting, that is for restricted or unrestricted availability of variables.

Additionally, we provide formulas for the model-averaging of different quantities of interest, such as net survival, crude probability of death and explained variation. The choice of quantity to be averaged may be driven by the main reported outcome. That choice can also be driven by the common denominator to all other measures (the individual excess hazard).

Furthermore, we combine models from the same class of models, given our choice of algorithm for model selection. The same strategy could be envisaged for averaging results from models arising from different model structure. We expand this point further in the discussion Chapter.

The work presented in this Chapter is offering many avenues for further research. Nonetheless, it is important to note that it is based on empirical evaluation of performances. As such, it relies on several strong assumptions, including:

(i) The non-parametric Pohar Perme estimator is considered to represent the true value of cancer survival. We recognise that this is a particularly strong assumption, however we acknowledge that the following statement holds true in practice: the Pohar Perme estimator is the estimator of choice when describing levels of survival in a given area. Therefore estimated levels of survival tend to be compared to the Pohar Perme estimator.

(ii) Excess hazard models estimate the same quantities as Pohar Perme, even when these are not using the same set of predictors.

We acknowledge that a series of simulations should complete the observations made here. Since with simulations, we would (1) know what the true distribution of future events is and how best to reproduce it, (2) be able to look at longer-term predictions, (3) be able to compare projections for cohort of patients further away from the training cohorts, (4) not need to assume (i) and (ii).

Discussion

This work was conducted in the context of persisting socio-economic inequalities in cancer survival in England despite inequalities being a public health priority within almost all recent policy initiatives for cancer diagnosis and treatment. Survival is a key metric for the evaluation of the performance of health system in relation to cancer. However, the very nature of survival means that there is a temporal gap between the data available and the cohorts of recently diagnosed patients. In order to offer more timely prognostic information about future cohorts of cancer patients, I aimed to evaluate and improve the methodology available for the prediction and projection of cancer survival.

Following their cancer diagnosis, patients receive information on the likely course of their disease from different sources:

i Clinicians who may provide a prognosis based on their experience and knowledge in addition to their intuition;

ii Patients themselves and their peers who might offer an evaluation of the course of their disease based on similarities with acquaintances, or case-stories;

iii Statistical models based on the influence of specific factors upon various outcomes including recurrence, disease progression or death, derived from cohorts of patients diagnosed in the past.

However different these sources are, the validity of their predictions relies on their ability to identify correctly recurring patterns and associations. In this thesis, excess hazard regression models for the prediction of cancer survival for populations of newly diagnosed patients have been described. I have explored two areas where additional research was needed, in order to develop the available methodology for the prediction and projection of cancer survival (i) model selection and (ii) model validation. The ideas developed within this thesis are novel and pave the way for future development, assessment, validation of and prediction from excess hazard regression models.

1 Excess hazard model selection

In Chapter 1, I focus on specifying the most appropriate functional forms of effects for key predictors of cancer survival. This is achieved through model selection algorithms. Given that linked cancer registration datasets gather large collection of patient-level information, there are many variables available and many complex associations which could potentially be modelled. Selecting the correct model is therefore challenging but crucial for correct inference.

I concentrate on two specific algorithms for model selection, both of which were developed in the context of multivariable model-building and time-to-event data. [88, 90] Their underlying philosophy differs in that they employ backward and forward selection strategies, but they also differ in terms of their views on simultaneous testing of non-linear and non-proportional effects. The structure of each model selection algorithm is set. Nonetheless the selection of effects is data driven.

Because survival patterns can be complex, detailed guidelines on model selection within the context of multi-variable excess hazard modelling has great potential benefit for health data analysts, epidemiologists and policy makers. In the population-based cancer survival setting, the number of potential explanatory variables remains low, but the shape of their effects may be complex, including non linearity and non proportionality. The selection of effects may be based on the literature, reflecting a range of expected and plausible effects. Algorithms are helpful tools to screen through the range of effects that one wishes to consider. Nonetheless, although rigorous algorithms for model selection are presented, there remains potential issues of model misspecification. These can arise for a number of reasons including: (i) assumptions made in the context of a specific model set are not verified in the data; (ii) outliers or clusters of data may distort the parameter estimates; (iii) important explanatory variables are not available; (iv) functional forms are misspecified; (v) missing information. Sensitivity analyses may be conducted to study the impact that unmeasured confounders, or missing information have on model specification. There are two potential extensions of my work which could widen its scope and use in practice.

1.1 Missing data

The first of these is the application of model-selection algorithms in the context of variables with missing information. The decision not to consider missing data is one of the main limitations of this work. Observational data often carry non-negligible proportions of missing information for key predictors. [168] In this thesis the analyses are restricted to cohorts of patients with least missing information on stage at diagnosis and performance

status ($< 15\%$). These relatively small proportions mean we could exclude the records with missing information from the analysis.

A common situation is where information is missing at random. That is the distribution of missing data can be described by the fully-observed variables available in the dataset, including the outcome of interest. In such instances, multiple imputation by chained equation (MICE) can be used to avoid biases and loss in efficiency that occur when using complete-case analysis. [169–172] MICE involves creating k completed copies of the datasets, with separate sets of imputed data for records with missing values. The imputation model must be compatible with the analysis ('substantive') model, for unbiased estimation of the regression coefficients. This is challenging when the substantive model is not yet defined at the time of imputation.

The excess hazard regression setting within which we perform model selection is at the crossroads between two active research areas in missing data methodology: (a) imputation when the effects are complex (non-linear), or involved in interactions; (b) non-standard distributions for the substantive model, such as the Cox model. Several methods and ideas for accounting for missing data in these settings have been proposed, tested and evaluated on real data. Some of these examine variable selection, [173] some the selection of functional forms of effects, [174, 175] and others time-dependent effects. [176] I list below some of their suggestions, with their limitations.

i Imputation of missing data assuming the most complex substantive model. [172]

Limitations of this approach are: (1) the most complex model may not be appropriate for the data, (2) all other substantive model tested after imputation need to be nested in the most complex model.

ii Variable selection done on the complete-case dataset only.

The main limit here is a lack of efficiency, and risk of bias since data are likely not missing completely at random.

iii Variable selection on a single imputation.

Estimated standard errors are too small since the imputed dataset is assumed to be observed, no account is taken of additional uncertainty.

iv Algorithm for variable selection applied to all k imputed datasets, combining their estimated parameters using the Rubin's rules and using the Wald test to test for the effects of each variable. [173]

This is a very computer intensive approach. The following simplification is offered: variable selection on all imputed data selecting either predictors that appear in any of the selected models, or only predictors that appear in all/half or more of the models.

v Variable selection on stacked imputed data.

This approach uses weights to correct standard errors that would otherwise be too small, due to the larger sample size.

Aside from multiple imputation, one could consider inverse probability weighting (IPW) as an alternative to deal with missing information. Here, each complete record would be weighted by the inverse of the probability of having no missing data based on their individual characteristics. These records would therefore be up- or down-weighted depending on the proportions of patients with similar characteristics and for whom missing data is present. A weighted model selection would then follow. Intuitively, this appears less disruptive to use such an approach to account for missing data in excess hazard model selection, and circumvent the issues raised above. However, there are two methodological considerations to bear in mind with IPW: (1) weighted regression methods are generally inefficient, leading to large confidence intervals, and (2) the uncertainty in the estimation of weights needs to be taken into account in the model selection.

To address the problem of missing data in excess hazard model selection, none of the directions discussed here can be readily implemented. As a first methodological step, one could compare the performance of a relatively simple suggestion for imputation (**ii** or **iii** above) to a more complex one or to IPW.

1.2 Synergy with penalized regression

When studying cancer survival, the minimum set of information includes the type of cancer, sex of the patient, the age at cancer diagnosis, and the follow-up time elapsed since cancer diagnosis. Traditionally, most analyses are stratified by cancer site and sex. For each combination, follow-up time and age are strong predictors of overall and cancer survival.

'Data simplification' is sometimes performed, when the number of explanatory variables is reasonable and relatively simple effects are expected, based on background knowledge or limitations of the data. [62–65, 96, 98] The main limitation of such an approach is potential mispecification. This may result from incorrect underlying assumptions about the effect of interest or as a consequence of other mispecified effects. [86] In order to minimise mispecification and maximise correct inference, I demonstrate in Chapter 1 the importance of a systematic model screening strategy being adopted alongside background knowledge, particularly when sample size allows. I provide useful practical guidance for such model selection, [91] bearing in mind that the more explanatory variables there are, the more complex models become and the higher the chances of mis-specification, if such careful model selection is not adopted.

Penalised regression splines stem from a very different philosophy to model building. It is in stark contrast with selection of functional forms of effects, since there is no selection *per se*.

The tensor product, [177] a multidimensional penalised spline function, was recently adapted to the context of cancer survival. [5, 56] The baseline hazard, non-linear and time-dependent effects, and interactions between continuous explanatory variables are modelled with flexibility via tensor products. These are products of unidimensional spline functions, such that the coefficient(s) for one variable are varying according to values of other variable(s). For some given dimension of the marginal spline bases, smoothing parameters need to be estimated rather than the functional forms. Even if one needs to specify the number of knots and their locations, these have less influence in a penalized than in an un-penalized context.

This strategy does not constrain the forms of continuous effects, and non-linearity, non-proportionality and interactions are specified all at once. The only constraints are the smoothing parameters (as many as there are variables) that control the flexibility of the estimated curve/plane. The method reaches its limits when the number of continuous explanatory variables exceeds four. This is due to the exponential growth in number of parameters. For example, if we modelled 3 variables with splines with 5 bases, the number of parameters to be estimated would be $5^3 = 125$ and with 4 or 5 variables, it would reach $5^4 = 625$ and $5^5 = 3,125$ parameters, respectively. To reduce the number of parameters, one strategy would be to select differing levels of complexity (dimension) for the splines. Variable selection becomes necessary for situations with over four predictors (including follow-up time).

Since tensor products and smoothing techniques lead naturally to prediction, [178] a further extension would be to study model-building algorithms in combination with using tensor products. All continuous predictors would benefit from being modelled with more flexibility, whether or not the data requires such flexibility. The smoothing parameter would then rectify over-parameterisation and avoid un-desirable over-fit. One could exploit the results of model-building strategies, ahead of using tensor products for inference: the degree of flexibility (dimension of the splines) in the tensor product regression could be based on the selected functional form of each predictor. For instance, continuous variables selected with a linear effect could be modelled with splines of lower dimension than those retained in the final model(s) with more complex non-linear effects.

I believe the strengths of the tensor product are the weaknesses of traditional regression models and vice versa. Borrowing from both sets of tools would allow better inference, especially for prediction, free from model mispecification.

2 Model validation: explained variation

The key achievement of Chapter 2 is the extension of a measure of explained variation to the context of net survival in the relative survival data setting, by ways of weights applied to all records. The weights reflect the probability that the observed event is an event of interest (here, death due to the cancer under study).

The biggest driver for the methodological developments presented in this thesis is the availability of large and richer datasets than was previously available. Among other predictors, it is now possible within UK data to measure the prevalence of comorbidities at the time of cancer diagnosis. [179] I will use the example of comorbidities as an example of application of the measure of explained variation. Presence of pre-existing comorbidities can significantly influence timely presentation for cancer diagnosis [180] and access to treatment, and ultimately survival. Understanding how the patterns of these comorbidities impact cancer survival is of great interest to clinicians and public health specialists, especially in an ageing multi-morbid population. Selection of the most relevant comorbidities, and their potential interactions require that a suitably simple strategy is in place. Given the large datasets and the potential challenge of multiple testing, distinguishing which comorbidities are most predictive of survival is likely to be much better achieved through measures of explained variation than current methods.

Chapter 2 also focusses on the machinery available for validating predictive models. There exist a large number of tools, mostly used for individual prediction models. [18–20] These can be divided into three main groups, based on what they assess: overall performance, calibration and discrimination. I review below two possible extensions of Chapter 2 together with their clinical and public health benefits.

Alongside the development of REw, I provide some thoughts on how the weights introduced could be used to adapt the Brier score and the ROC curve to the relative survival data setting. Combining these weights and time-varying ROC curves would enable the validation of markers derived from excess hazard models at different times after diagnosis. These future developments would be beneficial, especially in a context where medicine with targeted immunotherapy treatments for specific tumour and patient's characteristics become more widely available. Individual patients' predictions would become key in such a context and the validity of these predictions could only be assessed with standard calibration and discrimination measures adapted to the relative survival data setting.

Frailty models can also be fitted in the relative survival data setting. [57] These models account for individuals being clustered within health providers (region, hospital), and thus not independent anymore. A random effect that measures the between-cluster variability is

estimated. As yet, in the relative survival data setting, there is no specific tool that informs the analyst how important it is to take account of that variability. Explained variation could be a useful quantity to characterise the importance of accounting for frailty, as is currently advocated in 'standard' survival analysis. [181] Measuring REw in excess hazard models with random effects would be a further extension of this thesis. Comparing REw derived from a standard excess hazard model to REw derived from a frailty model would provide the proportion of variation explained by taking account of the clustered nature of the data.

3 Prediction of cancer survival: multi-model inference

Model-based predictions and projections are the focus of Chapter 3, in which I propose the implementation of multi-model inference. This approach is based on the belief that some models may receive equivalent support from the data, and selecting a unique one would be restrictive and possibly misleading for prediction. [157] Model selection is based on information criteria, therefore enabling inference to be made from models without using multiple testing.

The application of multi-model inference presented in Chapter 3 remains within the data culture [51] in which relationships between variables are simplified and specified through regression models, reliant on the data meeting the specific assumptions of the models. I use multi-model averaging, borrowed from the Bayesian framework [166] and the Information Theoretic approach, [157] to avoid relying on a single model for predictions. I present model-averaged individual excess hazard of death as it represents the natural output from the models. Besides, it is the common denominator for the derivation of other model outcomes. However, I also briefly consider averaging measures of net survival and crude probability of death as well as explained variation. This offers an advantage whereby the outcome of interest can be estimated from each selected model and averaged directly using weights (derived, for instance, from the AIC).

3.1 Ensemble learning

The methodology introduced in Chapter 3 is a first building block for addressing formally prediction and projection of cancer survival. In that Chapter, the focus is on model averaging within given iterative algorithms for model selection. However, one does not need to restrict oneself, and can also average the outcomes of models selected by different algorithms. Furthermore, although my regression models are all fitted on the logarithm of the excess hazard scale, other types of scales and models could be included and their AICs contrasted.

More widely, the philosophy of multi-model inference and the idea of using 'best' models from different algorithms bring us closer to the philosophy of ensemble learning. Among others, the stacked generalisation (or stacking) and the super learner are ensemble algorithms in which the predicted outcomes from classification schemes or any regression models are combined together, in such a way that their predictions minimise a pre-defined error. [182] In the machine learning field, one approach to making predictions, away from the regression models used in this work, would be with random survival forests. [183] This is an appealing ensemble learning method drawing from the random forest classification scheme, described by Breiman. [184] After adaptation to the relative survival setting, this approach could be used to predict cancer survival for patients diagnosed in newer cohorts, without making any parametric assumptions.

Such a variety of modelling approaches pull together increases predictive ability but simultaneously makes interpretation difficult or impossible. This is because the model parameters are not the key outcomes, but merely considered a nuisance. [185] In contrast, our approach aims to maintain interpretation while improving prediction. Interpretation is a feature that is key to this work, to help design public health interventions.

3.2 Enhanced survival predictions using dynamic variables

In all data analysed as part of this thesis, we access a rich collection of variables, especially for patients diagnosed in the most recent years. Nonetheless, one restriction consists in their values being only defined and available at the time of diagnosis. Looking at the time varying values of REw, we see that they decrease with follow-up time as the values of the predictors become less relevant to the current prognosis of surviving patients. Updating the status of variables such as stage, treatment, and performance status during follow-up would greatly benefit the predictive power of the models and the proportion of variation in outcome explained. Some updated information could be available from Hospital Episode Statistics (HES) datasets in which access and use of secondary care services are recorded for all patients at every hospital visit, pre- and post-diagnosis. This database is already the source of information on emergency admissions, [186] comorbidities, [179] and surgical treatment. [98] Updated treatment information could also be available from the Systemic Anticancer Therapy (SACT) data.

Possible models that accommodate dynamic variables include delayed entry models, Cox regression models with time-updated information, and the landmark prediction model. [187] Such a modelling structure allows estimations to be updated given the values of other covariates at any time before the estimation horizon. Estimations are performed on the patients for whom information is available. In the context of prediction for other cohorts of

patients, both frameworks would need to be extended to predictions outside of the study sample, or at horizons beyond what is observed for some patients.

4 Application: relevance for public health

This work is highly applicable to health services planning and provision and in policy implementation and assessment. Historically, the period approach has been used for the estimation of survival for patients for whom follow-up is not yet available. [188] This approach derives cancer survival in a similar fashion to life expectation and utilises only the survival experience of the patients alive in the most recent calendar period. [188, 189] Our approach uses multivariable regression models to predict cancer survival for cohorts of patients most recently diagnosed and utilises all the available data. Predictions of survival for an entire cohort, are valuable for public health purposes because they provide an early appreciation of the effects of public health interventions on survival. Such a preview is valuable for the continued assessment of the effects of interventions.

Ahead of their implementation, interventions could also be assessed based on fictional and simulated scenarios. Modelling fictitious changes in the incident patient population or in the effects of prognostic factors could enable one to foresee what might be likely outcomes of potential interventions that would have led to these changes. Such knowledge can help design interventions that are most likely to maximise patient benefit. This scenario-based approach is a logical next step to this work. One possible way of designing these scenarios follows the steps detailed below:

1 Based on the data available, models would be selected for the prediction of cancer survival, for patients most recently diagnosed, as described in this thesis.

2 Scenarios that may reflect changes in the general population (such as ageing) or effects of cancer-specific interventions (such as earlier diagnosis, access to surgery for elderly patients) would be designed.

3 Values of predictors would be modified for the patients diagnosed in the most recent cohort, according to these scenarios.

4 Step (3) would be repeated several times to account for uncertainty, leading to several scenario-based datasets.

5 The model(s)'s parameters estimated in (1) would be used to predict and project survival for the artificial cohorts obtained in (4).

For increased relevance to the public health context, the effects of variables reflecting the environment of the patients (hospital characteristics, conditions in which care is received, social support etc.) could be modelled, in addition to patient and tumour factors typically available from cancer registry data linked to other healthcare datasets. The benefits of some specific interventions may 'simply' be seen on survival, but could also be on complex contrasts. As an example, one might want to reduce inequalities in cancer survival. In this scenario it would be useful to make sure one knows what the baseline inequalities are, in order to assess progress against that benchmark.

In this thesis, the terms prediction and projection are used for a very specific context and purpose, that of a public health setting, in which policy and planning are being constantly assessed and monitored. Nonetheless, the ideas developed could certainly be beneficial to other settings: individual prognosis in the clinical context and causal inference in the epidemiological context, where predictions are a necessary preliminary step, ahead of the estimation of causal contrasts. [185, 190]

5 Conclusion

Health systems are complex networks which involve patient attitudes and behaviours, availability of care and social support, accessible equipment, streamlined communication between healthcare practitioners, efficient informatics system, so on and so forth. Unexpected side effects of health interventions can be seen in the system, although they originally aimed at improving health outcomes. The effects of such interventions should be routinely assessed, ideally before implementation, and then monitored to ensure uninterrupted efficiency. The methodology developed and proposed in this thesis adds to the tools available to make such assessment and monitoring possible when the outcome of interest is cancer survival.

Bibliography

[1] Bray F, Colombet M, Mery L, Piñeros M, Znaor A, Zanetti R, et al. Cancer Incidence in Five Continents, Vol. XI (IARC Scientific Publications No. 166). Lyon: International Agency for Research on Cancer; 2019.

[2] Belot A, Ndiaye A, Luque-Fernandez MA, Kipourou DK, Maringe C, Rubio FJ, et al. Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data. Clinical Epidemiology. 2019;11:53–65.

[3] Mozumder SI, Dickman PW, Rutherford MJ, Lambert PC. InterPreT cancer survival: A dynamic web interactive prediction cancer survival tool for health-care professionals and cancer epidemiologists. Cancer Epidemiology. 2018;56:46–52.

[4] Cutler SJ, Ederer F. Maximum utilisation of the life table method in analyzing survival. Journal of Chronic Diseases. 1958;8:699–712.

[5] Remontet L, Uhry Z, Bossard N, Iwaz J, Belot A, Danieli C, et al. Flexible and structured survival model for a simultaneous estimation of non-linear and non-proportional effects and complex interactions between continuous variables: Performance of this multidimensional penalized spline approach in net survival trend analysis. Statistical Methods in Medical Research. 2019;28(8):2368–2384.

[6] Giorgi R, Abrahamowicz M, Quantin C, Bolard P, Esteve J, Gouvernet J, et al. A relative survival regression model using B-spline functions to model non-proportional hazards. Statistics in Medicine. 2003;22(17):2767–84.

[7] Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. The Stata Journal. 2009;9:265–290.

[8] Bower H, Crowther MJ, Lambert PC. strcs: A command for fitting flexible parametric survival models on the log-hazard scale. The Stata Journal. 2016;16(4):989–1012.

[9] Cortese G, Scheike TH. Dynamic regression hazards models for relative survival. Statistics in Medicine. 2008;27(18):3563–3584.

[10] Expert Advisory Group on Cancer. A policy framework for commissioning cancer services (Calman-Hine report). Department of Health; 1995.

[11] Haward RA. The Calman-Hine report: a personal retrospective on the UK's first comprehensive policy on cancer services. The Lancet Oncology. 2006;7(4):336–46.

[12] Morris E, Haward RA, Gilthorpe MS, Craigs C, Forman D. The impact of the Calman–Hine report on the processes and outcomes of care for Yorkshire's colorectal cancer patients. British Journal of Cancer. 2006;95(8):979–985.

[13] Fleissig A, Jenkins V, Catt S, Fallowfield L. Multidisciplinary teams in cancer care: are they effective in the UK? The Lancet Oncology. 2006;7(11):935–43.

[14] Department of Health. The NHS Cancer Plan: a plan for investment, a plan for reform; 2000.

[15] Department of Health. Cancer Reform Strategy; 2007.

[16] Department of Health. Improving outcomes: a strategy for cancer; 2011.

[17] De Savigny D, Blanchet K, Adam T. Applied Systems Thinking for Health Systems Research: A Methodological Handbook. McGraw-Hill Education; 2017.

[18] Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. Bristish Medical Journal. 2013;346:e5595.

[19] Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Medicine. 2013;10(2):e1001380.

[20] Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Medicine. 2013;10(2):e1001381.

[21] Hingorani AD, van der Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. British Medical Journal. 2013;346:e5793.

[22] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of Internal Medicine. 2015;162(1):W1–73.

[23] Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine. 2019;170(1):51–58.

[24] World Health Organisation. International statistical classification of diseases and related health problems. Tenth revision. World Health Organization; 1994.

[25] Tyczynski JE, Démaret E, Parkin DM. Standards and guidelines for cancer registration in Europe. The ENCR recommendations. IARC Technical Publication No. 40. IARC; 2003.

[26] Commission on Cancer. Facility Oncology Registry Data Standards (FORDS), Revised for 2016. American College of Surgeons; 2016.

[27] Ellis L, Woods LM, Estève J, Eloranta S, Coleman MP, Rachet B. Cancer incidence, survival and mortality: Explaining the concepts. International Journal of Cancer. 2014;135(8):1774–1782.

[28] Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. European Journal of Cancer. 2009;45:747–755.

[29] Parkin DM, Chen VW, Ferlay J, Galceran J, Storm HH, Whelan SL. Comparability and quality control in cancer registration. IARC Technical Report No. 19. Lyon: IARC; 1994.

[30] Li R, Abela L, Moore J, Woods LM, Nur U, Rachet B, et al. Control of data quality for population-based cancer survival analysis. Cancer Epidemiology. 2014;38(3):314–320.

[31] Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association. 1958;53(282):457–481.

[32] Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society Series B (Methodological). 1972;34(2):187–220.

[33] Pohar Perme M, Stare J, Esteve J. On estimation in relative survival. Biometrics. 2012;68(1):113–20.

[34] Schaffar R, Rachet B, Belot A, Woods LM. Cause-specific or relative survival setting to estimate population-based net survival from cancer? An empirical evaluation using women diagnosed with breast cancer in Geneva between 1981 and 1991 and followed for 20 years after diagnosis. Cancer Epidemiology. 2015;39(3):465–72.

[35] Schaffar R, Rachet B, Belot A, Woods LM. Estimation of net survival for cancer patients: Relative survival setting more robust to some assumption violations than cause-specific setting, a sensitivity analysis on empirical data. European Journal of Cancer. 2017;72:78–83.

[36] Schaffar R, Rapiti E, Rachet B, Woods LM. Accuracy of cause of death data routinely recorded in a population-based cancer registry: impact on cause-specific survival and validation using the Geneva cancer registry. BMC Cancer. 2013;13(1):609.

[37] Woods LM, Rachet B, Riga M, Stone N, Shah A, Coleman MP. Geographical variation in life expectancy at birth in England and Wales is largely explained by deprivation. Journal of Epidemiology & Community Health. 2005;59(2):115–120.

[38] Antunes L, Mendonça D, Ribeiro AI, Maringe C, Rachet B. Deprivation-specific life tables using multivariable flexible modelling – trends from 2000–2002 to 2010–2012, Portugal. BMC Public Health. 2019;19(1):276.

[39] Morris M, Woods LM, Rachet B. A novel ecological methodology for constructing ethnic-majority life tables in the absence of individual ethnicity information. Journal of Epidemiology & Community Health. 2015;69(4):361–7.

[40] Maringe C, Li R, Mangtani P, Coleman MP, Rachet B. Cancer survival differences between South Asians and non-South Asians of England in 1986-2004, accounting for age at diagnosis and deprivation. British Journal of Cancer. 2015;113(1):173–81.

[41] Spika D, Bannon F, Bonaventure A, Woods LM, Harewood R, Carreira H, et al. Life tables for global surveillance of cancer survival (the CONCORD programme): data sources and methods. BMC Cancer. 2017;17(1):159.

[42] Rachet B, Maringe C, Woods LM, Ellis L, Spika D, Allemani C. Multivariable flexible modelling for estimating complete, smoothed life tables for sub-national populations. BMC Public Health. 2015;15:1240.

[43] Pohar Perme M, Esteve J, Rachet B. Analysing population-based cancer survival - settling the controversies. BMC Cancer. 2016;16(1):933.

[44] Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. International Journal of Epidemiology. 1999;28(5):964–74.

[45] Royston P. Flexible parametric alternatives to the Cox model. The Stata Journal. 2001;1(1):1–28.

[46] Royston P, Parmar MKB. Flexible parametric-hazards and proportional odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine. 2002;21:2175–2197.

[47] Royston P. Flexible parametric alternatives to the Cox model: update. The Stata Journal. 2004;4(1):98–101.

[48] Rubio FJ, Rachet B, Giorgi R, Maringe C, Belot A. On models for the estimation of the excess mortality hazard in case of insufficiently stratified life tables. Biostatistics. 2019. [Epub ahead of print].

[49] Danieli C, Remontet L, Bossard N, Roche L, Belot A. Estimating net survival: the importance of allowing for informative censoring. Statistics in Medicine. 2012;31(8):775–86.

[50] Shmueli G. To Explain or to Predict. Statistical Science. 2010;25(3):289–310.

[51] Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science. 2001;16(3).

[52] Polley EC. Super Learner. University of California, Berkeley; 2010.

[53] Belot A, Coleman MP, Magadi W, Kaur J, Peet M, Rowlands S, et al. Geographic patterns of cancer survival in England: Adults diagnosed 2003 to 2010 and followed up to 2015. Office for National Statistics; 2017.

[54] Exarchakou A, Rachet B, Belot A, Maringe C, Coleman MP. Impact of national cancer policies on cancer survival trends and socioeconomic inequalities in England, 1996-2013: population based study. British Medical Journal. 2018;360:k764.

[55] Remontet L, Bossard N, Belot A, Estève J. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. Statistics in Medicine. 2007;26(10):2214–2228.

[56] Fauvernier M, Roche L, Uhry Z, Tron L, Bossard N, Remontet L, et al. Multi-dimensional penalized hazard model with continuous covariates: applications for studying trends and social inequalities in cancer survival. Journal of the Royal Statistical Society Series C. 2019;68(5):1233–1257.

[57] Charvat H, Remontet L, Bossard N, Roche L, Dejardin O, Rachet B, et al. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. Statistics in Medicine. 2016;35(18):3066–3084.

[58] Esteve J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. Statistics in Medicine. 1990;9(5):529–38.

[59] Mahboubi A, Abrahamowicz M, Giorgi R, Binquet C, Bonithon-Kopp C, Quantin C. Flexible modeling of the effects of continuous prognostic factors in relative survival. Statistics in Medicine. 2011;30(12):1351–65.

[60] Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. Statistics in Medicine. 2007;26:5486–5498.

[61] Hills M, Rachet B, Falcaro M. strel2: a command for estimating excess hazard and relative survival in large population-based studies. The Stata Journal. 2014;14:176–190.

[62] Maringe C, Walters S, Butler J, Coleman MP, Hacker N, Hanna L, et al. Stage at diagnosis and ovarian cancer survival: evidence from the International Cancer Benchmarking Partnership. Gynecologic Oncology. 2012;127(1):75–82.

[63] Maringe C, Walters S, Rachet B, Butler J, Fields T, Finan P, et al. Stage at diagnosis and colorectal cancer survival in six high-income countries: a population-based study of patients diagnosed during 2000-2007. Acta Oncologica. 2013;52(5):919–32.

[64] Walters S, Maringe C, Butler J, Rachet B, Barrett-Lee P, Bergh J, et al. Breast cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK, 2000-2007: a population-based study. British Journal of Cancer. 2013;108(5):1195–208.

[65] Walters S, Maringe C, Coleman MP, Peake MD, Butler J, Young N, et al. Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004-2007. Thorax. 2013;68(6):551–64.

[66] Cowppli-Bony A, Uhry Z, Remontet L, Voirin N, Guizard AV, Tretarre B, et al. Survival of solid cancer patients in France, 1989-2013: a population-based study. European Journal of Cancer Prevention. 2017;26(6):461–468.

[67] Jooste V, Bouvier AM, Bossard N, Uhry Z, Coureau G, Remontet L, et al. Trends in probabilities of death owing to cancer and owing to other causes in patients with colon cancer. European Journal of Gastroenterology & Hepatology. 2019;31:570–576.

[68] Tron L, Belot A, Fauvernier M, Remontet L, Bossard N, Launay L, et al. Socioeconomic environment and disparities in cancer survival for 19 solid tumor sites: An

analysis of the French Network of Cancer Registries (FRANCIM) data. International Journal of Cancer. 2019;144(6):1262–1274.

[69] Rutherford MJ, Abel GA, Greenberg DC, Lambert PC, Lyratzopoulos G. The impact of eliminating age inequalities in stage at diagnosis on breast cancer survival for older women. British Journal of Cancer. 2015;112:S124–8.

[70] Syriopoulou E, Mozumder SI, Rutherford MJ, Lambert PC. Robustness of individual and marginal model-based estimates: A sensitivity analysis of flexible parametric models. Cancer Epidemiology. 2019;58:17–24.

[71] Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. Journal of Statistical Computation and Simulation. 2015;85(4):777–793.

[72] Bower H, Crowther MJ, Rutherford MJ, Andersson TML, Clements M, Liu XR, et al. Capturing simple and complex time-dependent effects using flexible parametric survival models: A simulation study. Communications in Statistics - Simulation and Computation. 2019:1–17.

[73] Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biometrical Journal. 2018;60(3):431–449.

[74] Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, et al. State-of-the-art in selection of variables and functional forms in multivariable analysis – outstanding issues; 2019. Available from: https://arxiv.org/abs/1907.00786.

[75] Harrell J F E, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine. 1996;15(4):361–87.

[76] Royston P, Sauerbrei W. Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables. Series in probability and statistics. John Wiley & Sons; 2008.

[77] Riley RD, Snell KI, Ensor J, Burke DL, Harrell J F E, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Statistics in Medicine. 2018;38(7):1276–1296.

[78] Schumacher M, Hollander N, Schwarzer G, Sauerbrei W. In: Crowley J, Ankerst D, editors. Prognostic factor studies. Chapman et Hall/CRC; 2006. p. 289–333.

[79] Austin PC, Allignol A, Fine JP. The number of primary events per variable affects estimation of the subdistribution hazard competing risks model. Journal of Clinical Epidemiology. 2017;83:75–84.

[80] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33(1):1–22.

[81] Kullback S, Leibler RA. On information and sufficiency. The Annals of Mathematical Statistics. 1951;22(1):79–86.

[82] Laud PW, Ibrahim JG. Predictive model selection. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):247–262.

[83] Greenland S. Modeling and variable selection in epidemiologic analysis. American Journal of Public Health. 1989;79(3):340–349.

[84] Efron B. Estimation and Accuracy After Model Selection. Journal of the American Statistical Association. 2014;109(507):991–1007.

[85] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer Series in Statistics. New York: Springer; 2009.

[86] Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. Statistics in Medicine. 2007;26(2):392–408.

[87] Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Statistics in Medicine. 2004;23(16):2509–25.

[88] Royston P, Sauerbrei W. Multivariable modeling with cubic regression splines: a principled approach. The Stata Journal. 2007;7:45–70.

[89] Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Statistics in Medicine. 2007;26(30):5512–28.

[90] Wynant W, Abrahamowicz M. Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. Statistics in Medicine. 2014;33(19):3318–37.

[91] Maringe C, Belot A, Rubio FJ, Rachet B. Comparison of model-building strategies for excess hazard regression models in the context of cancer epidemiology. BMC Medical Research Methodology. 2019;19(1):210.

[92] Sauerbrei W, Royston P, Look M. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. Biometrical Journal. 2007;49(3):453–73.

[93] Sauerbrei W, Royston P, Zapien K. Detecting an interaction between treatment and a continuous covariate: A comparison of two approaches [Journal Article]. Computational Statistics & Data Analysis. 2007;51(8):4054–4063.

[94] Wynant W, Abrahamowicz M. Flexible estimation of survival curves conditional on non-linear and time-dependent predictor effects. Statistics in Medicine. 2016;35(4):553–65.

[95] Butler J, Foot C, Bomb M, Hiom S, Coleman M, Bryant H, et al. The International Cancer Benchmarking Partnership: An international collaboration to inform cancer policy in Australia, Canada, Denmark, Norway, Sweden and the United Kingdom. Health Policy. 2013;112(1):148–155.

[96] Uhry Z, Bossard N, Remontet L, Iwaz J, Roche L. New insights into survival trend analyses in cancer population-based studies: the SUDCAN methodology. European Journal of Cancer Prevention. 2017;26:9–15.

[97] Quaresma M, Coleman MP, Rachet B. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971–2011: a population-based study. The Lancet. 2015;385(9974):1206–1218.

[98] Benitez Majano S, Di Girolamo C, Rachet B, Maringe C, Guren MG, Glimelius B, et al. Surgical treatment and survival from colorectal cancer in Denmark, England, Norway, and Sweden: a population-based study. The Lancet Oncology. 2019;20(1):74–87.

[99] Hudson JC. Elementary models for population growth and distribution analysis. Demography. 1970;7(3):361–8.

[100] Mickens RE. Mathematical and numerical comparisons of five single-population growth models. Journal of Biological Dynamics. 2016;10:95–103.

[101] Auger P, Poggiale JC. Emergence of population growth models: fast migration and slow growth. Journal of Theoretical Biology. 1996;182(2):99–108.

[102] Hisano M, Connolly SR, Robbins WD. Population growth rates of reef sharks with and without fishing on the great barrier reef: robust estimation with multiple models. PLoS One. 2011;6(9):e25028.

[103] Coale AJ, Demeny P, Vaughan B. Regional Model Life Tables and Stable Populations: Studies in Population. Elsevier; 2013.

[104] Manton KG, Land KC. Active life expectancy estimates for the US elderly population: a multidimensional continuous-mixture model of functional change applied to completed cohorts, 1982–1996. Demography. 2000;37(3):253–265.

[105] Sharp L, Black RJ, Muir CS, Gemmell I, Finlayson AR, Harkness EF. Will the Scottish Cancer Target for the year 2000 be met? The use of cancer registration and death records to predict future cancer incidence and mortality in Scotland. British Journal of Cancer. 1996;73(9):1115–21.

[106] Valipour AA, Mohammadian M, Ghafari M, Mohammadian-Hafshejani A. Predict the Future Incidence and Mortality of Breast Cancer in Iran from 2012-2035. Iran Journal of Public Health. 2017;46(4):579–580.

[107] Rodriguez-Manero M, Cordero A, Kreidieh O, Garcia-Acuna JM, Seijas J, Agra-Bermejo RM, et al. Proposal of a novel clinical score to predict heart failure incidence in long-term survivors of acute coronary syndromes. International Journal of Cardiology. 2017;249:301–307.

[108] Blondon M, Hugon-Rodin J. A clinical risk score to predict the incidence of postpartum venous thromboembolism. Evidence Based Medicine. 2017;22(3):98.

[109] Chambers JC, Satinder MK, Smith DD. How to choose the right forecasting technique. Harvard Business Review. 1971.

[110] Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. JAMA. 2017;318(14):1377–1384.

[111] Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. BMC Medicine. 2018;16(1):120.

[112] Wei LJ. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. Statistics in Medicine. 1992;11(14-15):1871–1879.

[113] Rubio FJ, Remontet L, Jewell NP, Belot A. On a general structure for hazard-based regression models: An application to population-based cancer research. Statistical Methods in Medical Research. 2019;28(8):2404–2417.

[114] Rahman MS, Ambler G, Choodari-Oskooei B, Omar RZ. Review and evaluation of performance measures for survival prediction models in external validation settings. BMC Medical Research Methodology. 2017;17(1):60.

[115] Graf E, Schumacher M. An investigation on measures of explained variation in survival analysis. The Statistician. 1995:497–507.

[116] Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine. 1999;18(17-18):2529–2545.

[117] Schumacher M, Graf E, Gerds T. How to assess prognostic models for survival data: a case study in oncology. Methods of Information in Medicine. 2003;42(5):564–71.

[118] Gerds TA, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biometrical Journal. 2006;48(6):1029–40.

[119] Schoop R, Graf E, Schumacher M. Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. Biometrics. 2008;64(2):603–10.

[120] van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. Statistics in Medicine. 2000;19(24):3401–15.

[121] Parkes CM. Accuracy of predictions of survival in later stages of cancer. British Medical Journal. 1972;2(5804):29–31.

[122] Henderson R, Jones M, Stare J. Accuracy of point predictions in survival analysis. Statistics in Medicine. 2001;20(20):3083–96.

[123] Porzelius C, Schumacher M, Binder H. A general, prediction error-based criterion for selecting model complexity for high-dimensional survival models. Statistics in Medicine. 2010;29(7-8):830–838.

[124] Blanche P, Dartigues JF, Jacqmin-Gadda H. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. Biometrical Journal. 2013;55(5):687–704.

[125] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–38.

[126] Steyerberg EW, Eijkemans MJ, Harrell J F E, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Medical Decision Making. 2001;21(1):45–56.

[127] Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98(9):683–690.

[128] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691–698.

[129] Schemper M, Stare J. Explained variation in survival analysis. Statistics in Medicine. 1996;15(19):1999–2012.

[130] Choodari-Oskooei B, Royston P, Parmar MK. A simulation study of predictive ability measures in a survival model II: explained randomness and predictive accuracy. Statistics in Medicine. 2012;31(23):2644–59.

[131] Korn EL, Simon R. Measures of explained variation for survival data. Statistics in Medicine. 1990;9(5):487–503.

[132] Henderson R. Problems and prediction in survival-data analysis. Statistics in Medicine. 1995;14(2):161–184.

[133] Sauerbrei W, Abrahamowicz M, Altman DG, le Cessie S, Carpenter J. STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. Statistics in Medicine. 2014;33(30):5413–32.

[134] Sauerbrei W, Collins GS, Huebner M, Walter SD, Cadarette SM, Abrahamowicz M. Guidance for the design and analysis of observational studies: The STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative. Medical Writing. 2017;26(3):17–21.

[135] Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review. 1950;78(1):1–3.

[136] Schoop R, Schumacher M, Graf E. Measures of prediction error for survival data with longitudinal covariates. Biometrical Journal. 2011;53(2):275–293.

[137] Schoop R, Beyersmann J, Schumacher M, Binder H. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. Biometrical Journal. 2011;53(1):88–112.

[138] Wu C, Li L. Quantifying and estimating the predictive accuracy for censored time-to-event data with competing risks. Statistics in Medicine. 2018;37(21):3106–3124.

[139] Kejzar N, Maucort-Boulch D, Stare J. A note on bias of measures of explained variation for survival data. Statistics in Medicine. 2016;35(6):877–82.

[140] Choodari-Oskooei B, Royston P, Parmar MKB. A simulation study of predictive ability measures in a survival model I: Explained variation measures. Statistics in Medicine. 2012;31:2627–2643.

[141] Choodari-Oskooei B, Royston P, Parmar MKB. A simulation study of predictive ability measures in a survival model II: explained randomness and predictive accuracy. Statistics in Medicine. 2012;31:2644–2659.

[142] Stare J, Perme MP, Henderson R. A measure of explained variation for event history data. Biometrics. 2011;67(3):750–9.

[143] Steyerberg EW. Clinical prediction models. Statistics for Biology and Health. Springer; 2019.

[144] Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Statistics in Medicine. 1990;9(11):1303–25.

[145] Blanche P, Gerds TA, Ekstrom CT. The Wally plot approach to assess the calibration of clinical prediction models. Lifetime Data Analysis. 2019;25(1):150–167.

[146] Lemeshow S, Hosmer Jr DW. A review of goodness of fit statistics for use in the development of logistic regression models. American Journal of Epidemiology. 1982;115(1):92–106.

[147] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EWO. Calibration: the Achilles heel of predictive analytics. BMC Medicine. 2019;17(1):230.

[148] Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000;56(2):337–44.

[149] Heagerty PJ, Zheng Y. Survival Model Predictive Accuracy and ROC Curves. Biometrics. 2005;61(1):92–105.

[150] Saha P, Heagerty PJ. Time-dependent predictive accuracy in the presence of competing risks. Biometrics. 2010;66(4):999–1011.

[151] Li L, Greene T, Hu B. A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. Statistical Methods in Medical Research. 2018;27(8):2264–2278.

[152] Blanche P, Latouche A, Viallon V. In: Time-dependent AUC with right-censored data: a survey. Springer; 2013. p. 239–251.

[153] Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in Medicine. 2004;23(13):2109–2123.

[154] Wolbers M, Blanche P, Koller MT, Witteman JCM, Gerds TA. Concordance for prognostic models with competing risks. Biostatistics. 2014;15(3):526–539.

[155] Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Statistics in Medicine. 2013;32(30):5381–97.

[156] Lorent M, Giral M, Foucher Y. Net time-dependent ROC curves: a solution for evaluating the accuracy of a marker to predict disease-related mortality. Statistics in Medicine. 2014;33(14):2379–2389.

[157] Burnham KP, Anderson DR. Model selection and multimodel inference. Second edition ed. Springer; 2002.

[158] Raftery AE. Choosing models for cross-classifications. American Sociological Review. 1986;51(1):145–146.

[159] Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle. Proceedings of the 2nd International Symposium on Information Theory. Budapest: Akademiai Kiado.; 1973.

[160] Schwarz G. Estimating the Dimension of a Model. Annals of Statistics. 1978;6(2):461–464.

[161] Breiman L. The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error. Journal of the American Statistical Association. 1992;87(419):738–754.

[162] Berk R, Brown L, Buja A, Zhang K, Zhao L. Valid post-selection inference. Annals of Statistics. 2013;41(2):802–837.

[163] Lee JD, Sun DL, Sun Y, Taylor JE. Exact post-selection inference, with application to the LASSO. Annals of Statistics. 2016;44(3):907–927.

[164] Efron B. Estimation and Accuracy After Model Selection. Journal of the American Statistical Association. 2014;109(507).

[165] Tibshirani RJ, Taylor J, Lockhart R, Tibshirani R. Exact Post-Selection Inference for Sequential Regression Procedures. Journal of the American Statistical Association. 2016;111(514):600–620.

[166] Hoeting J, Madigan D, Raftery A, Volinsky C. Bayesian Model Averaging: a Tutorial. Statistical Science. 1999;14(4):282–417.

[167] Raftery AE, Madigan D, Volinsky CT. Accounting for model uncertainty in survival analysis improves predictive performance. Bayesian Statistics. 1996;5:323–349.

[168] Di Girolamo C, Walters S, Benitez Majano S, Rachet B, Coleman MP, Njagi EN, et al. Characteristics of patients with missing information on stage: a population-based study of patients diagnosed with colon, lung or breast cancer in England in 2013. BMC Cancer. 2018;18(1).

[169] Rubin DB. Multiple imputation for nonresponse in surveys. John Wiley & Sons; 2004.

[170] Nur U, Shack LG, Rachet B, Carpenter JR, Coleman MP. Modelling relative survival in the presence of incomplete data: a tutorial. International Journal of Epidemiology. 2009;39(1):118–128.

[171] White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011;30(4):377–99.

[172] Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Statistical Methods in Medical Research. 2015;24(4):462–87.

[173] Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? Statistics in Medicine. 2008;27(17):3227–46.

[174] Morris TP, White IR, Carpenter JR, Stanworth SJ, Royston P. Combining fractional polynomial model building with multiple imputation. Statistics in Medicine. 2015;34(25):3298–317.

[175] Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC Medical Research Methodology. 2012;12:46.

[176] Keogh RH, Morris TP. Multiple imputation in Cox regression when there are time-varying effects of covariates. Statistics in Medicine. 2018;37(25):3661–3678.

[177] Wood SN. Generalized additive models: an introduction with R. Boca Raton: Chapman and Hall/CRC; 2017.

[178] Currie ID, Durban M, Eilers PH. Smoothing and forecasting mortality rates. Statistical modelling. 2004;4(4):279–298.

[179] Maringe C, Fowler H, Rachet B, Luque-Fernandez MA. Reproducibility, reliability and validity of population-based administrative health data for the assessment of cancer non-related comorbidities. PloS One. 2017;12(3):e0172814.

[180] Renzi C, Lyratzopoulos G, Hamilton W, Maringe C, Rachet B. Contrasting effects of comorbidities on emergency colon cancer diagnosis: a longitudinal data-linkage study in England. BMC Health Services Research. 2019;19(1):311.

[181] Gleiss A, Gnant M, Schemper M. Explained variation in shared frailty models. Statistics in Medicine. 2018;37(9):1482–1490.

[182] Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. European Journal of Epidemiology. 2018;33(5):459–464.

[183] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The Annals of Applied Statistics. 2008;2(3):841–860.

[184] Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32.

[185] Keil AP, Edwards JK. You are smarter than you think: (super) machine learning in context. European Journal of Epidemiology. 2018;33(5):437–440.

[186] Maringe C, Rachet B, Lyratzopoulos G, Rubio FJ. Persistent inequalities in unplanned hospitalisation among colon cancer patients across critical phases of their care pathway, England, 2011–13. British Journal of Cancer. 2018;119:551–557.

[187] van Houwelingen HC. Dynamic Prediction by Landmarking in Event History Analysis. Statistics in Medicine. 2007;34(1):70–85.

[188] Brenner H, Gefeller O. An alternative approach to monitoring cancer patient survival. Cancer. 1996;78(9):2004–10.

[189] Myklebust Tg, Aagnes B, Møller B. An empirical comparison of methods for predicting net survival. Cancer Epidemiology. 2016;42:133–139.

[190] Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference. International Journal of Epidemiology. 2019;[Epub ahead of print].